DETECTION AND DIAGNOSIS OF DISTRIBUTED DISTURBANCES IN CHEMICAL PROCESSES

by

Nina Frances Thornhill

A thesis submitted for the degree of Doctor of Philosophy of the University of London

Department of Electronic and Electrical Engineering
University College London
Torrington Place, London WC1E 7JE

January 2005

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

PhD THESIS

DETECTION AND DIAGNOSIS OF DISTRIBUTED DISTURBANCES IN CHEMICAL PROCESSES

Nina F. Thornhill

Abstract

This thesis has been submitted for the examination of PhD in the University of London. Its subject is the detection and diagnosis of distributed disturbances in chemical processes. A distributed disturbance affects many variables such as feed, product and recycle flows, column temperature and product composition. It may upset just a single unit for example a distillation column, it may be plant-wide if it affects a complete production process or even site-wide if utilities such as the steam system are involved. Disturbances have an impact on profitability because production and throughput may have to back away from their maximum settings to accommodate process variability.

The research has used signal processing, spectral analysis and non-linear time series analysis of measurements from routine process operations and has led to new applications of these methods in chemical process diagnosis. In particular, the use of principal component analysis on the power spectra of process measurements has given a breakthrough in the analysis of non-steady processes because the spectra are invariant to the lags and time delays that can make PCA unreliable in the time domain.

The thesis offers novel methods and theoretical insights to support the industrial activity of detection and diagnosis of distributed disturbances. A key insight has been that non-linearity in the time trends of plant measurements is greatest in those measurements closest to the root cause because mechanical filtering by the plant makes the signals more linear as the disturbance propagates away from the source. A non-linearity index derived from process measurements can therefore locate the root cause of a disturbance.

A feature of the work has been its focus on industrial implementation. The methods are demonstrated with data from real processes and care was taken to devise robust default settings of parameters in the algorithms to facilitate their application in unseen plants. As demonstrated in a case study, the outcomes of the work will significantly reduce the time spent on analysis and focus attention towards root causes of faults so that maintenance effort is directed effectively.

CONTENTS PAGE

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 OVERVIEW

1.1.1 Overview of the research project

Motivation: The title of the project is "Detection and diagnosis of distributed disturbances in chemical processes". Disturbances to process operation are unwanted deviations in the process variables such as temperature, pressure and flow rates that move the process away from its optimum settings and thus degrade product quality or reduce output. The issue of the detection and diagnosis of *distributed* or *plant-wide* disturbances has not received much attention previously in the process control or process systems literature although it has been recognized as a key issue facing the process industries (Qin, 1998; Desborough and Miller, 2002; Paulonis and Cox, 2003).

A disturbed plant generally operates less profitably than one running steadily because production and throughput may have to back away from their maximum settings to accommodate process variability. (Martin *et. al*, 1991; Shunta, 1995). Early detection and automated diagnosis will mean more profitable operation of large-scale chemicals and petroleum products manufacture.

A distributed disturbance is one which appears in several places in a process plant. It may upset a single unit such as a distillation column, it may be plant-wide if it affects a complete production process or even site-wide if utilities such as the steam system are involved. Automated methods are needed to detect a distributed disturbance. The challenge then is to trace the disturbance to its root cause because maintenance actions are more cost-effective if a root cause has been correctly identified. These tasks are the main objectives of the PhD.

The thesis addresses dynamic disturbances that persist over a time horizon of hours to days. A common example is a plant-wide oscillation. It does not address abrupt faults such as compressor trips leading to critical or abnormal situations which have been the focus of the Abnormal Situations Management consortium and others (see below). Nor does it consider slowly developing faults such as catalyst degradation or fouling of a heat exchanger. For these types of faults methods such as multivariate statistical process control (Section 2.4.2) have proved effective.

Research strategy and an example: The detection of distributed disturbances in process measurements requires their characterization with a signature that is distinctive of the disturbance. Examples of signatures are peaks in the power spectra or the pattern of zero

crossings in the autocovariance functions. Detection of a distributed disturbance is accomplished through the recognition of measurements having similar signatures. For the diagnosis of the root cause of the disturbance, signatures that become stronger closer to the root cause are required.

The concept is illustrated in Figure 1.1 from a BP-Amoco refinery (Kwinana, W. Australia) which shows a unit-wide disturbance in a distillation column. The time trend of the analyzer (top panels) shows the product composition leaving the top of column was varying in a undesirable way. The power spectrum of the time trend has a distinct spectral peak because of the oscillation. Similar oscillations and spectral peaks were present in the steam flow and temperature and the upset was therefore classified as a distributed disturbance. The steam flow measurement showed non-linearity, as seen by the non-sinusoidal nature of the oscillation and the presence of a harmonic in the spectrum. As will be explained in Chapter 6, the appearance of non-linearity in a time trend is a diagnostic signature for the root cause of the disturbance; in this case it was a faulty steam flow sensor.

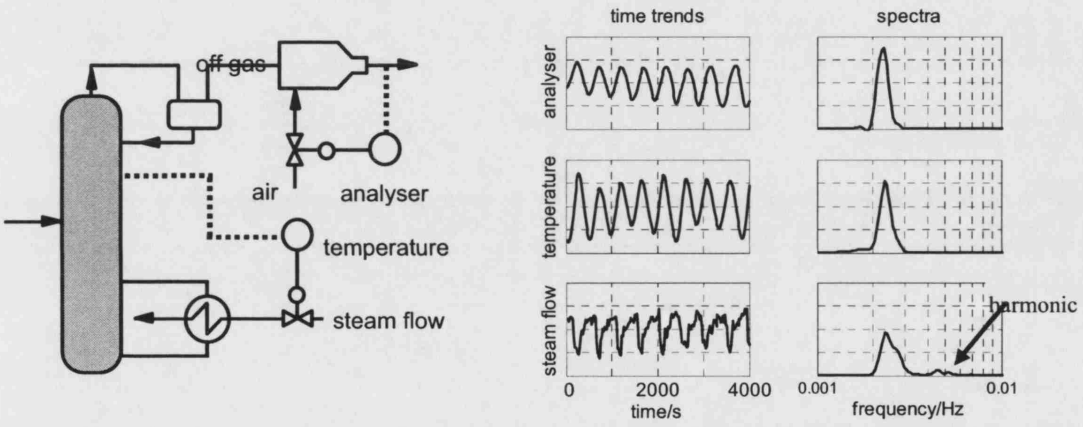

Fig 1.1 Illustration of a distrubuted disturbance in a distillation column.

Industrial relevance: Through the Abnormal Situations Management (ASM) consortium the National Institute of Standards and Technology, Honeywell and several operating companies have invested $16.6M in research and development towards elimination of abnormal situations and upsets in chemical plants. That level of funding suggests automated diagnosis of abnormal operations is a serious industry-wide concern.

The work of this thesis is thus relevant to petroleum and chemicals manufacturers because of their desire for increased automation. The work is not specific to any particular control system and therefore will be accessible to all process control equipment vendors. Attention has been paid to implementation issues and industrial examples as key steps on the way to successful technology transfer.

The industrial context: Industrial plant and controller performance monitoring systems and other asset management activities such as equipment condition monitoring are generally not run in real-time. They are part of the plant maintenance function rather than a real-time process operations function. Typically a plant audit is run on a fixed schedule or on demand, a report is produced and the engineering and maintenance staff then prioritize the problems and schedule the necessary maintenance.

By contrast, real-time operations are carried out by process operators at the control console. The operator's job is to drive the plant to optimise its profitability making use of tools such as optimizers and advanced controls. Operators deal with alarms which alert them to safely critical or economically critical situations. The vendor products supporting the operational activities are the distributed control systems (DCS) and supervisory control and data acquisition (SCADA) systems that provide the instrumentation and control functions to run the processes.

The research in this PhD provides a plant auditing and plant maintenance approach. Plant-wide disturbances represent a lower level of economic loss than a critical situation such as a plant shut-down. On the other hand, emergency shut-downs are rare occurrences whereas routine maintenance is always needed and indeed if it is done effectively there might never be an emergency shutdown.

1.1.2 Overview of the thesis

Main features of the research: The thesis describes original research undertaken and the results achieved during the PhD programme. The research has used signal processing, spectral analysis and non-linear time series analysis of measurements from routine process operations and has led to new applications of these methods in chemical process diagnosis. The methods are demonstrated throughout the thesis with data from real processes and care was taken to devise robust default settings of parameters in the algorithms to facilitate their application in unseen plants.

The thesis offers novel methods and theoretical insights to support the industrial activity of detection and diagnosis of distributed disturbances. As demonstrated in the case study of Chapter 8 the outcomes of the work will significantly reduce the time spent on analysis and focus attention towards the root causes of faults so that maintenance effort is spent effectively.

Content of the thesis: Chapter 2 of the thesis provides an assessment of current and past work in the field of disturbance detection and diagnosis in chemical process applications.

Chapters 3 to 6 describe the methods to be used with measurements from chemical processes and demonstrate their utility by applying them to an industrial data set. New software tools are presented that achieve detection and characterization of plant-wide disturbances and also give a means for the diagnosis of the root cause of one class of distributed disturbance, namely a non-

linearity such as valve friction that causes a limit cycle in a control loop which propagates to other measurements in a plant. The tools provide:

- High density plotting;
- Frequency domain filtering;
- Compression detection;
- Oscillation detection;
- Spectral principal component analysis (PCA);
- Non-linearity detection.

Implementation of the above techniques is illustrated in each chapter using an example from a hydrogen plant provided courtesy of a SE Asian refinery in which three plant-wide disturbances were present distributed amongst 37 plant measurements.

The software codes are based mainly on signal processing and data analysis methods that have previously appeared in text books or in the scientific literature. The novelty of the thesis lies in their application, for instance the work in spectral PCA is the first instance where PCA has been applied to the power spectra of process measurements. A powerful new way of visualizing PCA clusters in more than three dimensions is also presented. Some theoretical issues are addressed and resolved including a novel proof that determination of the period of an oscillation using zero crossings of the signal autocovariance function is more accurate than using the spectral peak in the frequency domain. The work in Chapters 3 to 6 has led to five refereed journal papers.

Chapter 7 investigates the physical basis for valve stiction which is a major cause of disturbance in chemical plants. The results from simulation of a first principles physics model give good insights into the origins of valve deadband and the phenomenon of stick-slip behaviour. A theoretical describing function analysis was completed jointly with Choudhury of the University of Alberta which indicates when and why limit cycles arise in control loop with sticking valves. A sixth journal paper that includes the work of Chapter 7 has been published.

Chapter 8 applies the methods to an industrial case study in collaboration with Paulonis and Cox of the Advanced Controls Technology group at the Eastman Chemical Company, Kingsport, Tennessee. The root cause of a plant-wide disturbance was found correctly on the first attempt. Measurements from the process before and after maintenance of the identified faulty valve are presented to demonstrate the improvement achieved.

Interdisciplinarity: The motivation for the use of signal processing and spectral methods arose because the author works in an electronic and electrical engineering department. The area of application, by contrast, is in the process control and monitoring area of chemical engineering. The thesis therefore draws upon both areas and is interdisciplinary.

1.2 NOVEL CONTRIBUTIONS

1.2.1 Contributions to dynamic principal component analysis

Chemical plants generally have many fewer modes of operation than measurements (Kresta *et. al.*, 1991). Thus many measurements tend to share co-ordinated or correlated behaviour. This observation has led to a huge body of work on plant-wide monitoring drawing on multivariate statistics for data-driven modelling and on-line fault detection using principal component analysis (PCA), and partial least squares (PLS). A drawback of the methods has been that they do not cope well with dynamic data because lags and time delays caused by process dynamics tend to destroy correlations. Therefore their impact on detection of disturbances such as persistent oscillation has been limited. These opinions will be backed up in the literature survey in Chapter 2.

A key breakthrough in this thesis has been to render PCA invariant to the time delays and lags caused by process dynamics. This has been achieved by means of integral transforms that remove the phase information of time delayed time trends. For instance, $\sin(\omega t)$ and $\cos(\omega t)$ are orthogonal functions with zero correlation in the time domain even though they are the same oscillation, merely out of phase by quarter of a cycle. Their power spectra, by contrast, have the same spectral peak and therefore are strongly correlated. Likewise, their autocovariance functions are insensitive to time delay and are identical. Chapter 3 utilizes the power spectra and autocovariance functions of plant data and thereby overcomes the problems of time delays that beset the multivariate analysis of dynamic data. The formulation of the methods and results from them have been published in a refereed journal paper (Thornhill *et. al.*, 2002).

1.2.2 Contributions to root cause diagnosis

A novel contribution in root cause diagnosis has been the application of methods from non-linear time series analysis (Kantz and Schreiber, 1997) to process data. The key insight has been that the non-linearity of the time trends of plant measurements is greatest closest to a non-linear root cause because the mechanical low-pass filtering effects of the plant make the signals more linear as they propagate away from the source. Early results from the PhD reported in a referred conference paper (Thornhill *et. al.*, 2000) gave an experimental demonstration of that finding. An industrial case study was described in Thornhill *et. al.* (2003*a*).

The thesis has an emphasis on technology transfer. Several sections of Chapter 6 were devoted to the discovery of robust optimized default parameters of the algorithm so that the methods can be rolled out easily to new and unseen plants. This detailed study and the recommended parameters has been published in IEEE Transactions on Control Systems Technology (Thornhill, 2005).

1.2.3 Industrial contributions of the thesis

The methods devised during the PhD have attracted industrial attention and several case studies have been completed in addition to those presented in this thesis. These include:

- Plant-wide disturbance analysis of the gas compressor train on a BP North Sea oil platform;
- Analysis of controller problems in Omuta Works, Mitsui Chemical Company, Japan;
- Analysis of rogue seasonality in the supply network of South Wales steel manufacturer, with Prof M. Naim of Cardiff Business School;
- Analysis of electrical power transmission system disturbances, with collaborators from Psymetrix, Edinburgh and Imperial College.

Successful applications with oil and chemical companies shows that the methods have achieved their aims of becoming industrially useful for process applications. UCL and technology vendor ABB are now engaged in a joint activity to turn the tools described in this thesis into a commercial product.

Applications in supply chain and electricity supply suggest that the methods are going to be transferable in future to other industrial sectors.

1.3 SPONSORS AND ACKNOWLEDGEMENTS

Work on the PhD started in 2001 and included sabbatical leave in Canada from January to September 2001 supported by the Department of Electronic and Electrical Engineering, University College London. The host was Professor Sirish Shah of the Chemical Process Control Group of the Department of Chemical and Materials Engineering at the University of Alberta. I am deeply grateful to Professor Shah for all he has done to encourage the work and for his sponsorship of the sabbatical as well as for his helpful and effective career guidance and mentoring over the past few years. The work was sponsored by the Royal Academy of Engineering (Foresight Award) and the NSERC-Matrikon-ASRA Industrial Research Chair in Process Control. The financial support of these sponsors is gratefully acknowledged.

I would like to thank Mick Flanagan at UCL who supervised the PhD and Chris Baker, the Postgraduate Tutor who has also read and commented on the work. They both have taken time out of their very busy schedules to help and give encouragement.

The collaborations with my University of Alberta co-authors Shoukat Choudhury, Biao Huang and Sirish Shah have been delightful and rewarding. Shoukat Choudhury was a PhD student at the time when the research of the thesis was conducted and we both benefited greatly from our detailed debates about the behaviour of processes, control loops and control valves.

I am grateful also to several industrial collaborators and co-authors and have learned much from them about the business of industrial control technology. Warren Mitchell, Dave Shook, Anand Vishnubhotla and Haitao Zhang of Matrikon (Edmonton, Alberta) made available data, computing facilities, office space and managerial time spent in discussing the work during the sabbatical. Similar thanks also to Keith Landells of BP Petrochemicals, Hull who contributed process insights relevant to the case studies.

I should also like to mention the excellent industrial collaboration offered by John Cox, Michael Paulonis and Jim Downs of the Advanced Controls Technology group at the Eastman Chemical Company, Kingsport, Tennessee. They have actively encouraged the work from the outset and provided data, their time and a willingness to conduct plant tests and to feed back the results. The opportunity for an academic researcher to become involved with plant tests and to compare data sets from before and after maintenance is rare and valuable.

Finally, in London I should like to say thank you to my colleagues in the Imperial College/UCL Centre for Process Systems Engineering, to its present and past directors and to the manger of its industrial consortium who have been providing intellectual inspiration and leadership ever since the Centre was established in 1989.

2. REVIEW OF PAST AND CURRENT WORK

2.1 OVERVIEW

2.1.1 Introduction

Chapter 2 of the thesis is a literature review. It maps out the field of study and places the work of the thesis in context. The focus of this PhD is data-driven analysis with an emphasis on methods that incorporate signal processing, for example through the use power spectra, autocovariance functions, signal filtering and non-linear prediction. The survey demonstrates that plant-wide disturbances can be detected and characterized by means of those techniques.

A taxonomy for fault detection and diagnosis literature is discussed first. It is the one proposed by Venkatasubramanian of Purdue University. The work of the thesis can mainly be classified as *quantitative process history based* analysis. An industrial survey follows giving a snapshot as of August 2004 of industrial applications and services available. The industrial section draws some conclusions about the types of technology and services being offered by vendors to end-users in different industrial areas. Lastly, the chapter goes into more technical detail in selected areas of relevance to the thesis and shows how the work of the thesis fits into context.

2.1.2 Themes of the review

The context of the thesis: The literature review will explain the reasons why a signal processing and data-driven approach was adopted in the project and why other approaches were not used. The main topics of the thesis are listed here:

- To use historical process data for plant auditing;
- To detect distributed disturbances that affect many measurements in a unit, plant or site;
- To determine which measurements are affected;
- To characterize the disturbances in order to gain insights into their severity and likely cause;
- To diagnose the root cause of each disturbance;
- To consider technology transfer issues.

Reference back to these requirements will be made as the review progresses.

Fault detection and diagnosis: Review papers in 2003 (Venkatasubramanian *et. al.*, 2003(a), 2003(b), 2003(c)) and a keynote paper by Venkatasubramanian at the IFAC CHEMFAS4 conference in June 2001 (Venkatasubramanian, 2001) made the economic argument for the importance of fault detection and diagnosis in the chemical process industries, quoting a figure of $20 billion lost each year due to incorrect responses to abnormal operating situations. The authors emphasized the critical industrial importance of abnormal event management (AEM) and thoroughly reviewed the contributions to AEM of fault detection, diagnosis and correction of process and equipment faults. They presented a family tree of methods for analysis of process systems with an emphasis on fault detection and diagnosis. Categories highlighted were:

- Quantitative and qualitative process history based methods;
- Quantitative and qualitative model-based methods.

The three part review has made a major contribution towards the academic maturity of the field of fault detection and diagnosis because for the first time all the disparate techniques have been brought together for discussion in one place. The strengths and weaknesses have been examined systematically and comprehensively and a hierarchical structure has been imposed that has had great benefits, for instance in showing that seemingly unrelated approaches have a common mathematical basis. A recent text book by Chiang *et. al.*, 2001 also present broad coverage of knowledge-based, data-driven and analytical techniques.

2.2 SURVEY OF FAULT DETECTION AND DIAGNOSIS

2.2.1 Introduction

The main classifications and areas identified in the reviews (Venkatasubramanian *et. al.*, 2003a, 2003b, 2003c) are outlined this section with a brief description of each and the citation of some key papers. Some additional practical issues not covered by the reviews are also addressed.

2.2.2 Model-based methods

Quantitative model based methods: Quantitative model based methods exploit a detailed model of a process for fault detection and isolation (FDI). The model is solved in real time subject to same inputs as those of the real plant and the values of the outputs are compared to those of the plant and conclusions are drawn when any mismatch is observed.

Quantitative model-based methods have a history of at least 20 years, it is a mature area and there are text books on the subject as well as comprehensive reviews. Early papers explaining the methods and setting out the foundations were Chow and Willsky (1984), reviews by Isermann (1984) and Basseville (1988), Gertler and Singer, (1990) and Frank (1990). Frank *et al,* (2000) reviewed some recent applications. Patton *et. al.* (2000) and Gertler (1998) are the main text books covering model based and residuals methods.

The models may be first principles models based on physics and chemistry, empirically derived models such as time series or a hybrid of the two. FDI seeks for inconsistencies between expected and observed behaviour and in a multivariable system the geometrical properties of the residual can indicate the source of a fault as well as its occurrence. The residuals vector is tested for randomness; any deterministic or non-random behaviour indicates the occurrence of a fault while the pattern of non-randomness among the various elements in the residuals vector helps to isolate and diagnose the fault. The analytical processing carried out with the model and residuals is based upon transformations which project measurements into a *feature space* where any faults would show up more prominently and be decoupled from one another. In general the nature of the expected faults has to be known in advance to do the appropriate transformations.

Model-based FDI based on first principles models is considered hard to achieve in the process industries because accurate, well calibrated and well validated models are expensive to create and maintain for processes which are often not well understood. The reaction chemistry and the thermodynamic properties of the materials being processed are the main source of uncertainty. Time series and other identified input-output models are more common.

The main industrial application at present for quantitative model-based methods is in condition monitoring of equipment and rotating machinery such as pumps and gas turbines but, as mentioned, quantitative model-based methods are not widely used in process applications. A mechanistic valve model used in Chapter 7 of this thesis illustrates some of the difficulties of the use of first principles models because it was hard to get the parameter values such as friction forces.

Qualitative model-based methods: Qualitative model based methods include techniques that capture fundamental causal relationships of a process based on physics and chemistry in a non-numerical way using classification and ranking. These methods were the subject of the second review paper by Venkatasubramanian *et. al.*, (2003b). Signed digraphs (SDG) are one of the main ways of representing causal qualitative knowledge. If a fault is observed then the model is used to deduce the root cause. Such causal models can be enhanced by the use of quantitative information concerning the magnitudes of influences, for instance that A influences B more than B influences A. Abstraction hierarchies decompose a system in terms of its subsystems, again for the purposes of diagnostic reasoning. An abstraction hierarchy called Multilevel Flow Modelling (MFM) has been championed by Petersen and Lind at the Technical University of Denmark (e.g. Petersen, 2000) as a way to describe the flows and storage of mass and energy at various levels and the causal relationships between them. The diagnostic strategies for searching causal

representations to find faults and fault propagation paths are reviewed in detail in Venkatasubramanian (2003(b)).

SDGs were extensively reviewed by Maurya *et. al.*, (2003a and 2003b). These papers discussed graph-based approaches for safety analysis and fault diagnosis of chemical process systems. They highlighted the difficulties of accurate capture of the graph representation pointing out that it is a time consuming task and error-prone. They showed how to develop graph models systematically and how to eliminate spurious and redundant branches and showed how to construct signed digraph (SDG) models for the difficult causalities associated with feedback control loops. Chen and Howell (2001) also considered fault detection and diagnosis using an SDG representation in the presence of feedback control. Early and influential papers which laid down the foundations were Årzén, (1994); Årzén *et. al.*, (1995), Kramer and Palowitch, (1987); Petti, *et. al.*, (1990); and Vedam and Venkatasubramanian, (1999).

Future work beyond the scope of this thesis and planned after the completion of the PhD, will aim for a linkage of data-driven process-history based methods and a qualitative model of the process derived from the process flowsheet such as an SDG. The aim will be to use numerical process history analysis to enhance an SDG by means of quantitative information about the magnitudes of influences.

2.2.3 Process history based methods

Introduction: The third paper in the three-part review by Venkatasubramanian *et. al.*, (2003c) discussed methods for fault diagnosis based on historical process knowledge, both historical data and also qualitative historical knowledge such as operators' expertise. They made comparisons in order to determine the areas of strength and weakness of each and concluded that all the methods have desirable features but that they need to be used together in a hybrid approach in order to exploit the strengths of each one. The things considered included, among others, were rapid detection and diagnosis; ability to isolate as well as detect faults; being able to identify new and unseen faults and ability to explain the reasons for decisions.

Qualitative process history based methods: Qualitative methods use process insights from operators or features described in natural language derived from past data. Examples include expert systems and qualitative process trend analysis.

The purpose of an expert system is to capture the knowledge of human experts and to perform the same sequence of actions as the human when the same conditions arise. The review by Venkatasubramanian *et. al.*, (2003c) is downbeat about use of expert systems for fault detection and diagnosis. Their opinion is that although the rules can be formulated systematically and transparently the scope can be limited and that the systems are difficult to maintain and update. However as will be seen shortly there are a number of commercially available solutions using

expert systems while an early, and successful expert systems project in the UK was published by Blue Circle Cement in a DTI funded study (Blue Circle Industries, 1990).

Qualitative process trend analysis describes the shapes and patterns of the time trends of process and forms conclusions about underlying faults from the shapes. An example is the characteristic triangular shape sometimes seen in the controller output of a flow loop with a deadband which was the focus of Rengaswamy and Venkatasubramanian (1995). Cheung and Stephanopoulos (1990) first described a template-based representation to extract features from data sets while Bakshi and Stephanopoulos (1994) used wavelet packets and other templates for multiscale representation of process data. The work provides a string of symbols that describe the process trend. Learned and Willsky (1995) have also used wavelets for feature extraction and were able to classify transient signals.

A qualitative process trend analysis was used in this thesis to detect piecewise linear compression in process data in Chapter 3 but otherwise the focus of the thesis has been on signal processing methods that characterise the time trends of measurements by means of spectra or autocovariance functions. In the author's opinion, the qualitative process trend analysis reflects a preference among chemical engineers to use the time domain. An influential text book in chemical process control led this approach starting in the 1960's (Shinskey, 1996) and it remains popular because the undergraduate chemical engineering curriculum does not have much coverage of frequency domain methods.

A more natural application domain for qualitative trend analysis is in batch processes to give a time domain description of a batch profile For instance in a fermentation it might be very significant if the lag phase at the start of the fermentation lasted longer than usual. An application for batch processes called DB-miner had multi-scale extraction of patterns, short term trends and local features in fermentation processes (Stephanopoulos *et.al.*, 1997).

Quantitative process history based methods: Process history based methods use information captured from previous running of the process. The quantitative methods are numerical and data-driven. The main distinction drawn by Venkatasubramanian *et. al.*, (2003c) is between principal component analysis (PCA) as a statistical classification method and neural networks as non-statistical classifiers.

The methods and literature of PCA and related multivariate statistical approaches are discussed in detail shortly. The main outcome of a PCA analysis is a sub-space with orthogonal axes which are the basis functions from which the original measurements may be reconstructed. As with the model-based residual methods, the benefit is that faults state are more distinct in the reduced space than in the measurements (Gertler *et. al.*, 1999).

The other main branch highlighted in Venkatasubramanian *et. al.*, (2003c) was neural network classifiers. They identified the papers describing the methods and there are also numerous applications. Poggio and Girosi (1990) first showed how neural network models are created from a non-linear combination of basis functions while Bakshi and Utojo (1999) have placed all the quantitative process history methods under a common framework of projection onto basis

functions, where the basis functions and type of project depend on the type of modelling method selected.

An issue in the analysis of process data using neural networks has been to include time histories, as discussed by Zhang (2001). Hybrid technologies have also been successful in which a multivariate statistical modelling method called PLS (projection onto latent structures) is combined with neural networks to accommodate non-linear relationships in input-output data (Qin and McAvoy, 1992; Baffi *et. al.*, 1999). Other successful hybrid approaches include Thompson and Kramer (1994) who showed how a hybrid system using the process knowledge of the operators and a neural network improved the performance in making predictions and in parameter estimation. Ignova *et. al.* (1996) used process knowledge derived from a sophisticated process model rather than from operator experiences while Glassey *et al.* (1997) discussed combining neural network modelling with an expert system for bioprocess supervision.

The work in this thesis did not use neural network technology. Neural networks provide powerful input-output models for the purpose of real-time predictions during a process operation but they are non-linear black-box models and it is not normally easy to drill down into their structure to uncover insights into the behaviour of the plant. They can be used to classify faults in a process if that fault has been encountered and analysed before. The thesis is focused towards plant auditing, an off-line method in which all available historical data are analysed together to extract features and signatures that reveal the locations and nature of new and unseen process faults. Neural network modelling was not considered the best choice for the application.

2.2.4 Six-sigma methods and statistical process control

Introduction: Six Sigma is the name given to a disciplined management approach to quality assurance that uses statistical visualization and problem-solving. The idea is to detect and understand the variables affecting quality and to eliminate the variability so that the number of defects or errors is effectively zero. The approach is very broad and is being used for both manufactured products and services. It is more prevalent in the manufacture of discrete items and in the service industries than in continuous chemical processes, nevertheless the terminology is being used by some technology vendors supplying the chemical process industries supplying controller performance monitoring tools to the chemical process industries and is reviewed here for that reason.

Six sigma grew from traditional univariate statistical process control (SPC) and statistical process monitoring (SPM). In univariate statistical process control the time trend of one variable is plotted on each chart. It is a well developed method whose history and techniques are documented in the text book by Wetherill and Brown (1991) and recently reviewed by Stoumbos *et. al.*, (2000) The academic focus in the process control community has moved away from univariate SPC and SPM methods and almost all the research involving process applications is

now in multivariate statistical process control and multivariate statistical monitoring. The sub-sections below review the univariate and multivariate literature.

Univariate SPC charts: Statistical process control is a real-time method in which measurements are processed as they become available and decisions are implemented immediately. In univariate statistical process control the time trend of one variable is plotted on each chart. The relevant British standard for construction of SPC charts is BS5700 which sets warning limits at 1.96σ and 3.09σ meaning that 95% of the measurements are within the warning limits and 99.8% are between the upper and lower alarm limits if the quality measurement has a Gaussian distribution. The aim of six sigma is to reduce variability by improving the process and tighten up the Gaussian distribution until the same alarm limit is at $\pm 6\sigma$, i.e. no more than about 3 per million of the quality measurements violate the alarm limits. The methods of this thesis can help by tackling the root causes of process variability.

The benefit of six sigma production in a process is to deliver reliable product while running as close as possible to the customer's specification. If the customer's specification is that the purity should be X then X defines the lower limit (-6σ). The upper limit represents an undesirable operating state, such as one where the purity and hence the cost is unacceptably high. A six sigma process virtually never violates the customers specification and virtually never enters the undesirable operating state. These concepts were introduced into the process industries in an influential book by Shunta of DuPont (1995).

The values to be plotted on the chart can be selected to suit the application. Often they are laboratory quality measurements but may also be derived quantities such as the average of the last three quality measurements or a measure of the spread between recent measurements.

The performance of an SPC chart is quantified in terms of the rate of false alarms or the average run length (Page, 1955; Champ and Woodall, 1987; Wetherill and Brown, 1991; Alwan *et. al.* 1994). False alarms during routine running are caused by the random occurrence of points lying at the extremes of the reference distribution. The charts should also be sensitive to disturbance, giving a true alarm soon after the onset of a disturbance. Mining of historical data identifies the disturbed episodes and gives a means of assessing how sensitive the selected chart limits are to true alarms.

Multivariate SPC: Multivariate SPC (MSPC) uses multivariate charts rather than univariate charts. The idea is that an upset to the correlations between variables can indicate a fault even if the variables themselves are all still within the limits of normal operation. The quantity plotted on a MSPC chart is abstract because it is usually a linear combination of several measured variables derived from projection of the current state of the plant onto a PCA model of normal in-control operation. For this reason MSPC charts are thought by some users to be difficult to interpret.

MSPC was placed in a broad context of quality and reliability engineering by Elsayed (2000). Key issues addressed were change point detection, correlated measurements, multivariate

systems and the integration of statistical control and process control. Goulding *et. al.* (2000) reviewed multivariate statistical process control methods in the process industries while Yoon and MacGregor (2000) gave an in-depth comparison between MSPC methods and the model-based fault detection methods mentioned earlier.

The concept of the warning and alarm limits is the same in the multivariate charts as in the univariate charts. It is necessary to find the 95% and 99.9% contours of the multivariable distribution by estimating the statistical distribution of the quantity being plotted in the charts which is not necessarily a multinomial distribution (Jackson and Mudholkar, 1979; Martin and Morris, 1996; Conlin et. al., 2000; Jia *et. al.* 2000; Westerhuis *et. al.*, 2000).

Relevance to the thesis: The thesis uses multivariate principal component analysis (PCA) in Chapter 4 but for historical audit purposes and not for a real-time MSPC application. The formulations of the two approaches are different, as will be explained shortly.

In particular the notion of confidence limits does not apply. When PCA is used for on-line monitoring then the warning and alarm limits are based on a reference data set from normal operation. The purpose of the analysis in this paper, by contrast, is the examination of historical data sets from periods of abnormal operation. The analysis is used to characterise the abnormalities and to highlight how many modes of abnormality are present by detecting clusters of measurements in the PCA plots.

2.2.5 Supporting methods for fault detection and diagnosis

Introduction: The topics outlined in this subsection are supporting methods that underpin quantitative process history based methods. Their relevance to the work of the thesis is assessed.

Gross error detection: Pre-processing of data can enhance fault detection, and anomalies detected in pre-processing may also give valuable insights for diagnosis (Pearson, 2001). Elimination of gross errors and some filtering are usually to be recommended. Famili *et. al.* (1997) discussed the topics of data visualization and multivariate methods as well as filtering and gross error removal. Their case studies showed the extent to which data pre-processing needs to be done manually. Turner *et. al.* (1996) also gave some guidelines about pre-processing to eliminate outliers, and noted that in their neural network modelling the data pre-processing step took up to 80% of the project time.

Bad data must always be removed. They arise because of transmission failure and appear in the data base as a BAD DATA text which gives errors when presented to a numerical algorithm, or as a physically infeasible number such as -999999. It is beneficial to remove these values and to replace them with estimates of the true value. Individual bad data points or small bad sections

may be replaced by linear interpolation between the good values either side but data sets with large amounts of bad data replacements should be discarded as unsuitable for analysis. Linear interpolation for bad data replacement can be detected in the same way as data compression so the comments about data compression in section 3.4 apply and the test described in section 3.4.5 can be used to decide when a data set is unsuitable for analysis.

Gross errors are values that are within the sensor range but wrong such as when a time trend captures a sensor calibration episode. Gross error removal can be based just on the properties of the data set or it can use some process knowledge, for example that an apparent change between one value and the next is not physically feasible. When statistical tests are applied it is usually to test whether an individual measurement is further than, say, three standard deviations from the mean. There are papers in the academic literature on this topic (Iordache *et. al.*, 1985; Jongenelen *et. al.*, 1988; Rollins and Davis, 1993; Narashiman and Mah, 1989; Pearson, 2001) as well as articles with a practical flavour in the trade magazines (Verneuil *et. al.*, 1992; Tham and Parr, 1994). The paper by Verneuil *et. al.* (1992) used process insights such as mass balances as well as statistical testing. Tham and Parr (1994) suggested ways in which bad data values can be replaced with estimated values. Tong and Crowe (1996) described the use of multivariate statistics through principal component analysis in gross error detection.

The work presented in this thesis has generally not applied any gross error removal. The justification for the approach is as follows:

- It is infeasible to spend 80% of the project time preprocessing the data. Methods that require such manual intervention are not good candidates for automation and technology transfer;
- Integral transforms such as power spectra and autocovariance functions are robust to gross errors in the time domain because the effects of a gross error (a delta function in the time domain) are spread evenly and thinly across the frequency domain;

The data can, however, be pre-processed by filtering. As will be seen in Chapter 3, filtering is beneficial for the removal of non-stationery trends such as diurnal variations while filtering in selected frequency ranges enables multiple oscillations to be detected.

Data reconciliation: Gross error removal is often combined with data reconciliation in which the gross error is replaced by a more realistic value that meets process constraints. In some cases, the whole data set may be adjusted to better fit a mass or energy balance. The data reconciliation literature is huge and was last reviewed by Crowe (1996). Albers (1994) gave a practical example of an application to a heat exchanger problem. A selection of academic papers that combine gross error detection with reconciliation is: Tjoa and Biegler (1991), Holly *et. al.* (1989), Crowe (1988).

Gross error detection and data reconciliation tools are commercially available. They are used in on-line optimization where plant data are aligned with a process model in order to estimate unknown parameters before running the optimizer (e.g. in AspenPlus On-Line from AspenTech). They also have uses in fault detection, for instance the detection of leaks that prevent a mass-

balance closure around a plant or unit. AspenAdvisor (AspenTech) is an example of a commercial tool for yield accounting and fiscal monitoring for accurate accounting of material entering and leaving a plant.

Data reconciliation was not used in the thesis because the dynamic features of the data were the focus of the work, not the steady state alignment of data with a process model (e.g. for closure of a mass balance). Indeed, the procedures generally use mean-centred data where the steady-state value has been disregarded.

Data compression: On-line data compression methods are implemented in the data historians of some commercial distributed control systems (DCS). The aim of data compression is to capture and store only those data points that represent the start and end of a trend (Mah *et. al.*, 1995; Hale and Sellars, 1981; Bristol, 1990).

Data are reconstructed after compression by interpolation between stored data points. The reconstructed data trends therefore consist of piecewise linear segments and are not the same as the originals, i.e. the compression is lossy. The use of compressed data in data-driven analysis of dynamic process data has been considered (Kudic and Thornhill, 1996; Watson *et. al.*, 1998). Those papers found compressed data to be of little or no value in any analysis where medium to high frequency dynamic effects are to be analyzed. With the cheaper costs of storage media equipment vendors are now placing less emphasis upon data compression. Honeywell's Loop Scout controller performance tool calls for uncompressed data, as does AspenTech's Performance Watch product.

Data compressed by an on-line method should not be used for the work of this thesis because high fidelity information about dynamic effects is destroyed by compression. Chapter 3 reports a technique for detection of the piecewise linear segments that characterize the industry-standard compression algorithms.

Change detection: Methods for detection of changes have been surveyed by Basseville (1988). These include signal processing methods (e.g. Basseville and Benveniste, 1983; Tugnait, 1982) whose purpose is to rapidly identify jumps in parameters in a dynamic time series model.

Papers on steady-state checking are also of interest for change detection, because a changing state implies the lack of a steady state. One approach converts the measurements into a qualitative description (i.e. "steady" or "not-steady") using statistical hypothesis testing. A null hypothesis is posed, for example that the measurements represent a steady state, and the measurements are tested to see if they have a distribution that can be accounted for by the null hypothesis. Narashiman *et. al.* (1988) have illustrated a method for determining if measurements have undergone a change of steady state using a parametric test based on the assumption of a Gaussian distribution.

Non-parametric statistical tests do not make any assumption about the probability distribution of the measurements and they may also be based on quantities that are not in engineering units. In a

non-parametric test the null hypothesis concerns, for instance, the number of changes of direction of data points in a trend. Kendall and Ord (1990) give a number of these statistical tests for detection of a trend. Koninckx (1988) applied the Wilcoxson test to detection of a steady state of a single variable. Another example of detecting process changes used the Kolmogorov-Smirnov test with data from a polyethylene process and from a North Sea pipeline (Thornhill *et. al,*. 1996*a*).

No automated change detection was used in this PhD project because the case studies were from steady state processes and were stationery. Nevertheless a recommendation of the work is to revisit the work of Thornhill *et. al,*. (1996*a*) to add a simple change detection routine to the toolbox of methods.

2.3 INDUSTRIAL APPLICATIONS

2.3.1 Introduction

This section reviews a selection of the commercial tools available for fault detection and diagnosis as of August 2004. It highlights the technologies offered by the vendors and draws conclusions about the business and technology drivers of various industrial sector that make one approach preferable to another.

The section has a focus on vendors supplying end users in the oil and chemicals industries. Other sectors such as electrical power and computer network management are also discussed selectively, however the vendor lists for those sectors are representative rather than comprehensive.

2.3.2 Oil and chemicals

ABB: ABB is a supplier of equipment and technology in the oil, gas and chemicals sector and also in power systems. Its catalogue includes control and laboratory instruments, control valves, control systems, machinery and drives and equipment for power supply industry such as transformers and switchgear.

The Industrial IT Extended Automation System 800xA plant automation system is their control system for batch and continuous process operations. It has a sub-system called Inform IT for integration of historical and process business data from controllers and equipment from which

key performance indexes (KPIs) can be determined. Optimize IT provides asset management of plant equipment to aid efficient plant maintenance. The applications aid data storage, retrieval and reporting and have a data base structure that captures the relationships between plant equipments so that, for instance, any equipment contributing to a low KPI in one area of the plant can be detected.

ABB's main application of quantitative model-based simulation is in training simulators. Training simulators tend to be very detailed requiring every valve and actuator to be modelled as well as the responses of the process. The ABB's STS-2000 training simulator can also be used for engineering analyses formulated by the user and it interfaces with the DCS as if it were the real plant.

ABB will be providing a product called LPM in 2005 for control loop performance assessment and tuning. LPM will include a module called PDA (Plant Disturbance Analysis) based on the techniques reported in this thesis.

AspenTech: AspenTech sells applications that address management systems, work practices and operations. Aspen Operations Manager-Event Management helps with putting events in context and to direct actions according to a standard operating procedures. The events in question are both technical such as process alarms and businesses events such as loss of a customer.

Aspen Watch is the product which does controller performance monitoring. It provides tools based principally on signal processing methods that detect problems and present information in a way that makes it easy for an engineer to investigate problems. It has a focus towards advanced controls and also monitors the active constraints and assesses lost economic opportunities when constraints are not pushed or when the wrong constraints are active.

Aspentech has great strength in all the technical sub-sections of section 2.2. For instance Aspen advisor uses knowledge based methods and rigorous modelling for data reconciliation while IOModel and Apollo provide input-output and first principles modelling capabilities. Aspen IQ provides tools for inferential sensing. It is, however, left to the user or other software suppliers to apply the powerful AspenTech tools to process- or equipment- specific applications.

Expertune: Expertune is an independent software vendor. They have a product called Plant Triage which does process performance monitoring and diagnosis and a controller tuning software called PID Tuner. Plant Triage has a focus on the regulatory and advanced control layers and its performance reports are cast in terms of which control loops and actuators need attention in order to enhance the economic performance of the plant. It uses the improvement methods of six-sigma to guide maintenance activity.

Honeywell: The basic distributed control system (DCS) supplied by Honeywell is called TPS, and recent enhancements are TDC3000 which provide advanced application such as a data historian and advanced multivariable control. TPS and TDC3000 are being integrated within the

Process Knowledge System (PKS) which does process, business and asset management. The aim is that anyone in a PKS site can access relevant information about a process from the dollars per hour profit to the maintenance status of an individual compressor. As with other DCS vendors specific applications such as fault detection and diagnosis are left to the user or to specialist suppliers.

The Honeywell Abnormal Situation Management consortium was set up in the late 1990's to map the way for Honeywell to address major plant failures. To date it has resulted in a new Honeywell approach towards alarm performance assessment rather than in new products. The Honeywell web site says the key outcomes of the ASM project have been to define areas in need of improvement and to understand how to achieve those improvements. These areas were organizational and cultural and include management systems, work practices and operations.

Honeywell's response to the findings of the ASM project has been to offer alarm performance assessment as a *service* which requires clients to collect alarm system data from their site and send it to Honeywell. The service then analyses the data and send the client a report which gives benchmarking against industry standards, detailing where alarms are not correctly configured, gives suggested reasons for nuisance alarms and also an evaluation of operators' responses to alarms. The technical parts of the analysis give advice such as identification of tags which generate large quantity of alarms due to process, configuration, or system causes, and also patterns of alarm configuration which contribute to alarm flooding. The aim is to help companies to move towards well-engineered alarm management systems in order to minimize abnormal situations in the plant.

The service approach has worked well for Honeywell in the area of controller performance assessment (their Loop Scout service). A finding of this thesis, however, is that there is a limit to what can be achieved in diagnosis without a knowledge of the process and equipment layout, a comment which applies to Loop Scout and may apply also to the alarm performance service. Honeywell's service approach detaches the analysis from process understanding rather than integrating them and it is up to the client to develop model-based or process history based methods to use with the results so that they can decide upon appropriate actions.

Invensys Foxboro: Foxboro has the A2 DCS system for batch and continuous processes. It has a focus on accessibility of process data so that users can develop applications using a package called Wonderware. Wonderware has a univariate SPC module called QI Analyst for real-time charting and also for analysis of historical data. The DT Analyst module monitors faults and efficiencies of process equipment. Much of the needed information is captured automatically, for instance the length of time needed to process materials from different suppliers while the reasons for downtimes are input manually by operators in a systematic way. The DCS system and Wonderware provide data collection, archiving and reporting that make it possible for users to develop applications of their choice such as fault detection and diagnosis. Wonderware does not provide the applications themselves and instead offers independent software vendors the

opportunity to sell their applications through Invensys. Companies that have been acquired by Invensys also supply some of these solutions. Examples are:

- DATACON from Invensys Simsci-Esscor for data reconciliation, gross error detection and location;
- DYNSIM from Invensys Simsci-Esscor for first principles process simulation with rigorous thermodynamic and fluid flow calculations, used for plant design and control strategy validation.
- ARPM from Invensys Simsci-Esscor that combines rigorous simulation with data reconciliation and parameter estimation for applications such as monitoring of compressor efficiency, monitoring catalyst activity, evaluating heat exchanger fouling and the potential benefits of cleaning.
- FACTNET from Pacific Simulation which is a multivariate analysis system together with a pattern recognition system that users can train with their own process expertise, soft sensing and sensor validation.

Matrikon: Matrikon is a mid-sized software and consulting company. Its major business is in software interfaces to provide connectivity between all the various industry-standard interfaces. It also has process and controller diagnosis software products. Their ProcessGuard software does alarm management for reporting and reviewing of alarms and event incidents, elimination of nuisance alarms, storage and analysis of alarms. The aim is to help operators make sense of alarms from the control systems, to learn from past events and to respond appropriately.

ProcessMonitor is a non-invasive online fault detection and diagnosis software that uses multivariate statistical methods. It captures the key relationships between variables and has visual outputs that help operators see and quickly correct for emerging abnormal operating problems before they cause alarms.

Matrikon's ProcessDoctor is a control loop performance monitor that detects problems with individual control loops and with advanced multivariable control structures and aids the user with data-driven signal processing such as oscillation and spectral analysis to diagnose the problems. These products interface with existing DCS systems and provide some of the fault detection and identification (FDI) applications made possible by the central collection and storage of process data by the DCS.

PAS: PAS (Plant Automation Systems) has applications called ControlWizard for real-time performance management and TuneWizard for tuning and optimization of process control assets with a focus on PID controllers, sensors and valves. The analyses are based on quantitative process history methods including signal processing, for instance they have sticking valve detection. The economic impacts are assessed to highlight the priorities for maintenance. Like Expertune and Matrikon they recommend the six-sigma approach to improvement and provide the tools to document progress.

They also have an alarm management suite which has helps maintain and audit plant alarms and to ensure that alarms that have been turned off for some purpose do not get forgotten. A Performance Metrics Manager application also provides key performance indicators for asset management and business monitoring (e.g. running costs). The software gives a structure to help users collect data and to do the calculations that turn the data into useful information and also to drill down to find the parts of the operation that contribute to poor performance.

Comments: The DCS vendors (ABB, Honeywell and Invensys) are meeting the end-users' requirements for systematic collection, storage and organization of data. It is assumed that the clients will supply their own fault detection and diagnosis applications or purchase them from independent software companies. A future trend especially in large and economically important plants such as in ethylene crackers is towards model-based control and optimization and some vendors are supplying the modelling capability. Opportunities will therefore exist in future to exploit the linkage of process models and analysis of process data for cause-and-effect reasoning. The models that are in use at the present time are not yet being made available for general purpose applications such as FDI, however.

The independent companies (Matrikon, Expertune and PAS) are supplying applications for plant performance assessment including controller performance. They are using quantitative process history-based methods such as multivariate statistics and signal processing. The six-sigma approach is used by these independent software companies as a structured way forward to prioritize maintenance and enhance process and controller performance.

The work of the thesis is closest to the controller performance assessment applications (PlantTriage from Expertune, ControlWizard from PAS and ProcessDoctor from Matrikon). The author's previous experience in single loop controller assessment using minimum variance benchmarking was the starting point of the work and it was that experience which motivated the plant-wide approach. The plant-wide approach is not yet used by any of the vendors, although PlantTriage and Process Doctor do have the capability for detection of multiple oscillations.

2.3.3 Equipment, rotating machinery and power systems

Emerson Process Management: Emerson Process Management's AMS suite provides asset management and condition monitoring for turbo, rotating and other mechanical equipment, electrical systems, process equipment such as pumps, heat exchangers and boilers, and also for instruments, and valves. Their fault detection and diagnosis of turbines and rotating machinery is based on rigorous thermodynamic models and for other equipment a variety of methods are mentioned at their web site involving special measurements used specifically for fault finding such as vibration and infrared thermography. The emphasis is on predictive maintenance.

Impact Technologies: Impact Technologies supplies advanced machinery diagnostics for aircraft, power, nuclear and the defence industries. They use signal processing techniques such as digital filtering, signal correlation and FFT and also model-based techniques based on empirical engine models for both the steady-state and transient operating regimes. They detect faults using a matching method. In the case of mechanical engine faults the diagnosis is mainly based on vibration analysis with these spectra of the vibration frequencies and structural resonances held in a knowledge base and rule based systems are used to match observations onto faults held in the data base.

GE Energy. GE Energy has products for monitoring power plants called GateCycle and EfficiencyMap. The approach is of model building allowing users to create plant models with gas turbines and other relevant equipments selected from a library. The purpose is to run the plant profitably and also to measure and track plant performance changes.

The models are hybrid combining first principles heat balance and machine efficiency curves customized for each manufacturers' machine. They have numerous applications including predictive capability such as examination of different operating modes, pressure losses and component fouling. The equipment models can also predict the expected performance of major equipment and monitor degradation between expected and current performance. This is an example of model-based fault diagnosis.

The major equipments involved are turbines, compressors, expanders and combustors. As with the Emerson approach, the success of model-based methods in the GE Energy products reflects the reality that certain types of equipment are (a) easier to model accurately and (b) economically worthwhile to model accurately because there are very many users of the same equipment.

Psymetrix: Psymetrix is a company based in Edinburgh that provides applications for improvement of the operation of power systems and to diagnose the causes of oscillating power flows in the network which sometimes reach a state from which normal operation is not recoverable. Major problems of thus type have been observed in major blackouts in the USA and Europe. They use signal processing and input-output model based methods to understand and identify the interactions between groups of generators in an electricity network.

The application of fault detection and diagnosis in power systems is going to grow in future because of new measurement technology with accurate time-stamping so that measurements can be reliably synchronized from separate locations. This will lead to better data for both the quantitative model-based and history-based methods.

Comments: When fault diagnosis is applied to specific equipments it is termed *condition monitoring* and the overall activity is called *asset management.* Emerson and GE Energy are explicitly using first principles models in fault diagnosis systems for equipment such as rotating machinery, engines and gas turbines and Invensys Simsci-Esscor are also using first principles

models in their ARPM application for monitoring of compressors. In these cases it is the vendors who create the models, in contrast to AspenTech who assumes the end users or other consultants will create the models. The reason for this difference is probably that there are many similar compressors, turbines and other machines in the field and a few well engineered models can be profitably rolled out to multiple customers.

By contrast, it is considered not cost effective to use first principles model based methods in chemical processes except for the largest and most economically important. However, the barriers to model building are coming down because of the growth of computing power and model libraries. The use of accurate first principles models in chemical processes will grow in future. At present, however, process history based methods are accessible to most end users and do not cost much because the measurements are readily available and already being used for other purposes such as closed loop control. Therefore, for now, applications such as multivariate statistical process control and signal processing methods for performance assessment are in the ascendancy.

2.3.4 Other industries

Gensym Corporation: Gensym is a company that provides an industrially leading expert system platform called G2 that is used for control, optimization and diagnosis of technical systems ranging from process plants to battle fields. They provide the reasoning engines and the client provides the applications with Gensym consultants assisting in the specification and implementation. An expert system application from Dow Chemicals presented at Gensym User Group Meeting, Boston, April 9-11 2003 optimized the usage of energy system equipment and also monitored equipment behaviour. This is an example of a qualitative process history-based method.

Gensym also has a neural network product called NeurOn-Line that fits the quantitative process history-based category. The neural network models are based on historical process data that capture the relationships between product quality and process conditions. Minnovex, a Canadian mining company, reported an application in ore grinding at a remote location in Papua New Guinea where the aim was to match the mill power to a vary variable feed tonnage and density. One attractive feature of the G2 solutions was said to be ease of maintenance in the remote location.

The industrial feature that characterizes the users of Gensym products is their very complex processes The Dow example had to take account of the impact of variable energy contracts in the operation and in the mining example the feedstock was very variable and the process dynamics were not well known. Although the given examples were control applications rather than fault detection and diagnosis their success suggests expert systems and neural networks may be the

solution of choice for complicated process run by heuristic rules or based on empirical correlation.

SUN Systems and Wild Packets: A growth area for fault analysis and diagnosis systems is in computer networks. A company called Wild Packets has a fault analysis system called Omni 3 which uses expert system technology together with data filtering and event logging to determine unusual activity in a computer network. The Sun Management Center 3.5 for Solaris on SUN computers helps computer network managers with in-depth diagnostics of hardware and software. It also uses a knowledge base of suggested actions in response to alarms on Sun systems. Predictive failure analysis warns of developing problems before they become critical.

2.4 DISTURBANCE DETECTION

2.4.1 Introduction

The thesis uses quantitative process history based methods to achieve detection and diagnosis of distributed disturbances. This section of the Chapter 2 places the methods in a broader context of related work in the area of chemical process control and gives some background. Each chapter of the thesis will also include a focused review of the literature that is directly relevant to the methods being presented in that chapter.
The next sub-section sets out the formulation of principal component analysis (PCA) as it is used for real-time multivariate statistical process control and then lays out an alternative formulation used in this thesis that is appropriate for process auditing. Spectral PCA, one of the main contributions of this PhD, is developed from that starting point. Section 2.4.5 reviews the techniques of oscillation detection in chemical processes.

2.4.2 Principal component analysis

Introduction: The methods of PCA or Principal Component Analysis have been widely applied for the detection of correlated measurements such as would be expected in the presence of a plant-wide disturbance (Kresta *et. al*, 1991, Wise and Gallagher, 1996; Goulding *et. al.*, 2000). Descriptions of the methods of PCA may be found from many sources, for example in Chatfield and Collins, (1980), Wold *et. al.*, (1987). Kourti and MacGregor (1996) wrote an overview and commented on the practical issues for industrial implementation.

PCA formulation for MSPC: PCA is derived from the singular value decomposition of the data matrix X. In a conventional multivariate SPC application (MSPC) the columns of X are the time tends of m measurements each sampled at N time instants where in general $N > m$. The columns are usually mean centred and scaled to unit standard deviation.

$$\begin{matrix} & m \text{ measurements} \rightarrow \\ \boldsymbol{X} = & \begin{pmatrix} x_1(t_1) & \cdots & x_m(t_1) \\ \cdot\cdot & \cdot\cdot & \cdot\cdot \\ x_1(t_N) & \cdots & x_m(t_N) \end{pmatrix} \begin{matrix} N \text{ time} \\ \text{samples} \\ \downarrow \end{matrix} \end{matrix}$$

The singular value decomposition is $\boldsymbol{X} = \boldsymbol{UDV'}$ where matrix $\boldsymbol{U}$ is N-by-m, $\boldsymbol{V'}$ is m-by-m, and $\boldsymbol{D}$ is diagonal and its elements are the positive square roots of eigenvalues of the m-by-m matrix $\boldsymbol{X'X}$. The principal component decomposition is expressed as $\boldsymbol{X} = \boldsymbol{TW'}$, where $\boldsymbol{T} = \boldsymbol{UD}$ and $\boldsymbol{W'} = \boldsymbol{V'}$.

A description of the majority of the variation in $\boldsymbol{X}$ can often be achieved by truncating the PCA description. The following is a three-PC model in which the variation of $\boldsymbol{X}$ that is not captured by the first three principal components appears in an error matrix $\boldsymbol{E}$:

$$\boldsymbol{X} = \begin{pmatrix} t_{1,1} \\ \cdots \\ t_{N,1} \end{pmatrix} w'_1 + \begin{pmatrix} t_{1,2} \\ \cdots \\ t_{N,2} \end{pmatrix} w'_2 + \begin{pmatrix} t_{1,3} \\ \cdots \\ t_{N,3} \end{pmatrix} w'_3 + \boldsymbol{E}$$

The $w'-$vectors are orthonormal row vectors with m elements and are called *loadings*. The $t-$vectors are column vectors with N elements and are called *scores*. The above formulation is the basis for multivariate statistical process monitoring because the $w'-$vectors capture typical profiles of sensor values. For instance in a plant with three sensors $w'_1 = (0.848 \quad 0 \quad -0.53)$ would mean there is a typical behaviour of the plant which can be described verbally as:

Sensor 1 is significantly higher than average, and sensor two is average,
and sensor 3 is below average.

If measured plant data had that same profile at, say, time step 223 then score $t_{223,1}$ would be large and $t_{223,2}$ and $t_{223,3}$ would be zero. That is to say, the scores at time step 223 indicate that the w'_1- profile is the best match to the current set of sensor measurements and the profiles captured in the w'_2- or w'_3- vectors are not good matches at that time (they may match better at other times, however).

Practical computations: Applications are available in the market for multivariate analysis. They include: MSPC+ (MDC Technology Ltd, Teeside, UK, now part of Emerson Process Management), PLS_Toolbox 3.0 (Eigenvector Research, Manson, WA, USA), ProcessMonitor

(Matrikon, Edmonton, Canada), SIMCA-P (Umetrics AB, Umeå, Sweden), and Unscrambler (CAMO ASA, Oslo, Norway).

PCA formulation for plant auditing: In an alternative formulation the data matrix is transposed so that the rows are the time trends of the measurements:

$$\begin{matrix} & N \text{ time samples} \rightarrow \\ \boldsymbol{X} = & \begin{pmatrix} x_1(t_1) & \dots & x_1(t_N) \\ .. & .. & .. \\ x_m(t_1) & \dots & x_m(t_N) \end{pmatrix} \begin{matrix} m \\ \text{measurements} \\ \downarrow \end{matrix} \end{matrix}$$

The $\boldsymbol{w}'-$ vectors now have N elements and resemble time trends. In this formulation PCA is a function approximation method in which each time trend in a row of $\boldsymbol{X}$ is built up from a linear combination of the $\boldsymbol{w}'-$ vectors. The $\boldsymbol{w}'-$ vectors are a set of orthogonal basis functions for the observed time trends. Similarities in the scores may then be used to detect clusters of similar time trends.

Such an approach has potential for plant-wide disturbance detection because all the measurements affected by a disturbance would have similar time trends.

Time shift PCA: It is known that time shifts in plant data caused by delays and lags in the process dynamics need special attention. The method of time shifted PCA has been used (Ku *et. al.*, 1995; Lakshminarayanan *et. al.*, 1997) in which the columns of the data matrix representing measurements with time delays are either time shifted, or the matrix is augmented with one or more time shifted replicates of the vector. Wise and Gallagher (1996) give a summary of the method. Another approach to alignment of data trends used dynamic time warping (Gollmer and Posten, 1996, Kassidas *et. al.*, 1998). Kassidas *et. al.* used that method with PCA for recognition of faults having different time duration.

Time shifted PCA has been extended beyond the requirements of fault recognition. Time shifting can identify dynamic relationships in the form of a time series model (Wise and Ricker, 1992; Ku *et. al.*, 1995) or a dynamic model using multivariate modelling of input-output data (Lakshminarayanan *et. al.*, 1997; Chen *et. al.*, 1998). Multiple time-shifted copies of data trends are included in the data matrix and a PCA analysis determines which of the time trends are linearly related. In that manner relationships in the form of a time series model may be established, for instance:

$$y_1[i] = ay_1[i-3] + bx_1[i-1]$$

where y_1 is a measurement in the plant and x_1 is an input variable or some other plant measurement that is causally related to y_1. The data matrix becomes very large when an arbitrary number of lagged time trends augment the data matrix. Shi and MacGregor (2000)

reviewed techniques for multivariate dynamic process modelling and gave a comparison of numerical methods for reduction of the large lagged variables matrix.

Disadvantages of time shifting: Time shifting may be problematical for several reasons (Thornhill *et. al.*, 2002). The true delay may not be a whole number of sampling intervals. For instance, if the delay were 3.2 minutes and the sampling interval were 1 minute then the closest alignment achievable would be a shift of 3.0 minutes and the time trends would remain out of alignment by 0.2 min. Secondly, there may be no unique time shift that achieves alignment. Oscillations at two unrelated frequencies may originate in control loops A and B and propagate plant-wide. Propagation delays from A to a given measurement point are not the same as from B because the routes through the plant are different and therefore it is not normally possible to find a unique time shift that aligns both oscillations simultaneously.

In the latter case, and also if time delays are unknown, it is necessary to include multiple time-shifted time trends in the data matrix. The data matrix then becomes very large and will generate more principal components than necessary. That makes it hard to reliably detect clusters of related measurements. These examples show that time delays pose problems in the detection of plant-wide oscillations and other types of time-varying disturbances. An aim of this thesis is to develop methods for multivariate analysis which are invariant to time delays and phase lags.

2.4.3 Spectral principal component analysis

Spectral analysis of processes Many plant-wide disturbances, especially those of an oscillatory nature, are not localized in time and thus the Fourier transform and power spectrum provide a natural means for their analysis. This section describes prior work that has utilized spectra in disturbance detection and diagnosis.

It is necessary to find all the measurements or control loops in a plant having the same disturbance because the root cause will be among that group. Pryor (1982) highlighted the usefulness of autocovariance functions and spectra for such a purpose. Desborough and Harris (1992) used the power spectrum to conclude that control loops had a long-term deviation from set point, and also to highlight an oscillatory loop, while Tyler and Morari (1996) have demonstrated a spectral signature arising from a disturbance in data from a Shell refinery. Thornhill and Hägglund (1997) showed that harmonics in a spectrum enabled loop tuning faults to be distinguished from faults due to non-linearity such as valve friction leading to a limit cycle.

Early plant-wide assessments, done manually, were used by Harris *et. al.* (1996) who reported plant-wide control loop assessment in a paper mill which they found spectral analysis of the univariate trends was useful. Ruel and Gerry (1998) found the non-linear root cause of an oscillatory plant-wide disturbance in a paper mill through an analysis of the harmonics of the

power spectra of the measurements. The root cause was a circulating pump in a tank which periodically failed to pump as pulp consistency varied.

PCA using spectra: Near infrared spectra (NIR) and infrared spectra are routinely analyzed by PCA or used in Partial Least Squares (PLS) models for estimation of analyte concentrations in unknown samples (Karstang and Henrikson, 1992; Riley *et. al.*, 1997; Yeung *et. al.*, 1999), and Seasholtz (1999) described the impact of multivariate analysis of NIR and other spectroscopy at Dow Chemical. In those cases the instruments themselves created the spectra, but the spectra may also be calculated numerically. Multivariate analysis of the spectra calculated from acoustic signals was reported by Belchamber and Collins (1993) who gave examples of the classification of the acoustic spectra in 24 frequency channels from a pump and an industrial blender. Spectral principal components analysis has been successfully used for acoustic monitoring applications in vehicles (Wu *et. al.*, 1999) and machinery (Tan *et. al.* 2002).

Spectral PCA of dynamic data: The novel feature of spectral PCA (Thornhill *et. al.*, 2002) is that the rows of the data matrix, $\boldsymbol{X}$, are the single-sided power spectra $P(f)$ of the time trends of the plant measurements, thus:

$$\boldsymbol{X} = \begin{matrix} N/2+1 \textit{ frequency channels} \rightarrow \\ \begin{pmatrix} P_1(f_0) & \cdots & P_1(f_{N/2}) \\ .. & .. & .. \\ P_m(f_0) & \cdots & P_m(f_{N/2}) \end{pmatrix} \end{matrix} \begin{matrix} \\ \textit{spectra of } m \\ \textit{measurements} \\ \downarrow \end{matrix}$$

The outcome of spectral PCA is the detection of clusters of plant measurements having similar spectra, which are interpreted to mean they are influenced by a common disturbance. The arrangement of the matrix means that the $\boldsymbol{w}'-$ basis functions are spectrum-like vectors (i.e. they have a frequency axis). The formulation given is therefore a functional approximation in which the spectra in the rows of $\boldsymbol{X}$ are reconstructed from the basis functions. As mentioned earlier, this interpretation of PCA lends itself to plant auditing.

Multivariate spectral analysis has several advantages over analysis in the time domain. It gives improved signal to noise ratio if the spectral content of the wanted signal is narrow-band compared to the noise. The power spectrum is invariant to time delays or phase shifts caused by process dynamics. It is also insensitive to missing values and outliers because the transforms of such effects are spread thinly across all frequencies in the spectrum. Bad data represented as large magnitude values such as –999999 should, however, be removed because the transform of such a large number would not only spread across the spectrum but would dominate it.

PCA using other integral transforms: The use of autocovariance functions in the data matrix has been shown to give similar results to spectral PCA (Tan *et. al.*, 2002; Thornhill *et. al.*, 2002). Other reported uses of PCA with transforms of the data matrix have focussed on multiscale wavelet transforms for on-line multivariate statistical process monitoring. Bakshi and

Stephanopoulos (1994) and Bakshi (1998) used wavelets and other templates for multiscale representation of process data in combination with PCA for modelling of data having features that changed over time and frequency. Kosanovich and Piovoso, (1997), Luo *et. al.*, (1999) and Shao *et. al.*, (1999) have also had success with multiscale PCA using wavelets. Wavelet functions are localized, however, so the wavelet transform is not invariant to time delays.

2.4.4 Other matrix factorizations

Independent component analysis: Independent Component Analysis (ICA) is a linear transform of a data matrix that minimises statistical dependence between the basis vectors. Its benefit is that it provides basis functions with a better one-to-one relationship with the physical sources of signals. Application of ICA to the sounds of a cocktail party is supposed to identify the individual speakers in the room.

Its foundation was explained by Comon (1994) and a recent special issue on blind source separation in the International Journal of Adaptive Control and Signal Processing (Davies, 2003) shows that ICA is starting to attract attention for control applications.

PCA ensures that the $\boldsymbol{w}'-$ vectors are orthogonal whereas ICA the transformation makes the vectors independent. The data matrix is represented as:

$$\begin{matrix} & N\,time\,samples \rightarrow \\ \boldsymbol{X} = & \begin{pmatrix} x_1(t_1) & \cdots & x_1(t_N) \\ \cdots & \cdots & \cdots \\ x_m(t_1) & \cdots & x_m(t_N) \end{pmatrix} \begin{matrix} m \\ measurements \\ \downarrow \end{matrix} \end{matrix}$$

The PCA decomposition is $\boldsymbol{X} = \boldsymbol{TW}' + \boldsymbol{E}$ and the rows of $\boldsymbol{W}'$ are orthogonal:

$$\sum_{\ell=1}^{N} w_i'(\ell) w_j'(\ell) = 0 \text{, for } i \neq j$$

where $w_i'(\ell)$ and $w_j'(\ell)$ are the ℓ'th elements in row i and j respectively of $\boldsymbol{W}'$.

The ICA decomposition is $\boldsymbol{X} = \boldsymbol{AS}' + \boldsymbol{F}$ where the rows of $\boldsymbol{S}'$ are independent and $\boldsymbol{F}$ is an error matrix representing the difference between the ICA reconstruction and the data matrix $\boldsymbol{X}$. Independence means that, if s_i is an element of row vector $\boldsymbol{s}_i'$ and s_j is an element of row vector $\boldsymbol{s}_j'$, then

$$\Pr(s_i, s_j) = \Pr(s_i)\Pr(s_j)$$

where $\Pr(s_i)$ and $\Pr(s_j)$ are the probability density functions of s_i and s_j and $\Pr(s_i, s_j)$ is the joint probability density function. Hyvärinen and Oja (2000) showed that statistical independence can be approximated when the kurtoses (the fourth moments) of the distributions

$\Pr(s_i)$ of the row vectors are maximised and gave an algorithm called FastICA for combining the elements in the $\boldsymbol{X}$ matrix in a way that generates an $\boldsymbol{S}'$ matrix whose rows have probability density functions with maximised kurtosis. The definition of independence treats s_i' and s_j' as random vectors with values s_i and s_j. In reality there is just one data set (one realization of the random processes) and the probability density functions are estimated from the histograms of the values of the elements in each row while statistical moments such as kurtosis are calculated from the values as:

$$kurt(y) = \frac{\frac{1}{N}\sum_{k=1}^{N} y^4 - 3\left(\frac{1}{N}\sum_{k=1}^{N} y^2\right)^2}{\left(\frac{1}{N}\sum_{k=1}^{N} y^2\right)^2}$$

where the $y-$values are the elements in a vector such as a row of $\boldsymbol{S}'$.

Li and Wang (2002) gave the first example of application of ICA to the analysis of dynamical chemical process trends and dimension reduction. Kano *et. al.*, (2003) and also Lee *et. al.* (2004*a*) used ICA instead of PCA in an on-line multivariate statistical process monitoring application with good results while Lee et al (2004*b*) explored the use of ICA for dynamic data using the time lagging method.

Spectral ICA: ICA can also be applied to the power spectra of process data. Xia and Howell (2003) and the PhD thesis of Xia (2003) gave the first application of ICA to process spectra. A demonstration of what can be achieved with ICA as a post-processing step following a spectral PCA analysis is demonstrated in Xia *et. al.*, (2005) who showed that ICA produces a very useful set of spectral-like basis functions with just one spectral peak in each independent component. The task of diagnosis of the sources of oscillations in a chemical plant was greatly aided by enhanced correspondence between independent components and plant oscillations.

Non-negative matrix factorization (NMF): NMF is a method related to PCA and ICA. It seeks for an alternative set of basis functions that enable a better one-to-one correspondence between physical sources and basis functions. A key paper in *Nature* by Lee and Seung (1999) addressed the issue of face recognition from two dimensional images. They said in their paper: "The basis images for PCA are *eigenfaces* some of which resemble distorted versions of whole faces. The NMF basis is radically different: its images are localized features that correspond better with intuitive notions of the parts of faces."

The matrix decompositions are:

$\boldsymbol{X} = \boldsymbol{TP}' + \boldsymbol{E}$	$\boldsymbol{X} = \boldsymbol{AS}' + \boldsymbol{F}$	$\boldsymbol{X} = \boldsymbol{WH}' + \boldsymbol{G}$
PCA	*ICA*	*NMF*

where in PCA the rows of $\boldsymbol{P}'$ are orthonormal. In ICA the independence property discussed earlier $\Pr(s_i, s_j) = \Pr(s_i)\Pr(s_j)$ holds, while in NMF the elements of both $\boldsymbol{W}$ and $\boldsymbol{H}'$ are non-negative. Matrices $\boldsymbol{E}$, $\boldsymbol{F}$ and $\boldsymbol{G}$ are the errors in reconstruction when a reduced number of basis functions are used.

Clearly NMF is applicable only when $\boldsymbol{X}$ is non-negative, as would be true if the rows of $\boldsymbol{X}$ are power spectra. The relationship between ICA and NMF for non-negative $\boldsymbol{X}$ has been established by Plumbley (2002) who showed the ICA decomposition can be transformed to an NMF decomposition by a matrix rotation. A recently submitted paper from the University of Alberta (Tangirala and Shah, 2005) is the first report of the use of NMF to power spectra in the area of process fault detection and diagnosis.

Spectral ICA and NMF as extensions to the work of the thesis: Spectral ICA and NMF are proving useful additions to the basic method of spectral PCA. The benefit of ICA as a post-processing step to spectral PCA was recognized by Chunming Xia of the University of Glasgow who selected is as a topic for his PhD (Xia, 2003). The author of this thesis was introduced to the method of NMF by Arun Tangirala of the University of Alberta in the summer of 2004 (Tangirala and Shah, 2005). Both ICA and NMF have the benefit that spectrum-like basis functions only have positive peaks, unlike in spectral PCA where the basis functions have both positive-going peaks and negative-going valleys. Moreover, the results so far have shown that when multiple oscillations are present then each spectral peak tends to appear in a separate basis function. Xia *et. al.* (2005) have successfully exploited this feature in disturbance diagnosis.

2.4.5 Oscillation detection

Oscillation detection using zero crossings: Oscillations are a common type of plant-wide disturbance and their regularity permits some specialized techniques for their detection. Several authors have addressed the detection of oscillatory measurements in process data. Hägglund (1995) detected zero crossings of the error signal in a control loop and calculated the integrated absolute error (*IAE*) between successive zero crossings. An oscillatory time trend has larger *IAE* values than a random one. Hägglund's paper described a real-time method and its incorporation into the ECA400 autotuning PID control module from Alfa Laval Automation. The control module generated an alarm when oscillations were detected.

Thornhill and Hägglund (1997) inspected the regularity of the zero crossings. The presence of an oscillation was inferred when the regularity was high and an estimate given of the signal to noise ratio of the oscillation. The zero crossing detection method had to explicitly address the problem of noise in the time trend, however because it required a method for elimination of the spurious zero crossings caused by fluctuating noise in the vicinity of the true zero crossing point. The

method of Forsman and Statin (1999) also used zero crossings and *IAE*. By distinguishing between positive and negative deviations they were able to detect asymmetrical oscillations.

Use of autocovariance functions: The horizontal axis in an autocovariance function (*ACF*) plot is time-like, representing the lag ℓ between the time trend and a delayed version of the time trend. The vertical axis is the correlation between the time trend and the delayed trend. An estimator for calculation of the *ACF* from a data set is:

$$ACF(\ell) = \frac{1}{N-\ell-1} \sum_{i=\ell+1}^{N} \tilde{x}(i)\tilde{x}(i-\ell)$$

where $\tilde{x}$ is the mean centred data scaled to unit standard deviation.

The *ACF* of an oscillating signal is itself oscillatory with the same period as the oscillation in the time trend. Moreover, the averaging provided by the summation in the above expression means that the *ACF* is much less noisy than the time trends. Thus it is beneficial to use the *ACF* for oscillation detection. For example, if the absolute value of the *ACF* at the first minimum exceeds a threshold then the possibility of an oscillation is inferred. The patented method of Miao and Seborg (1999) utilized additional cycles of the oscillatory autocovariance function. The authors demonstrated that the method could distinguish a decaying oscillation from a sustained oscillation.

The detection of oscillatory plant-wide disturbances using zero-crossings of autocovariance functions has been demonstrated in this thesis. The technique will be described in Chapter 4.

2.5 DISTURBANCE DIAGNOSIS

2.5.1 Introduction

Root causes: The focus in the previous section was the detection of plant-wide disturbances. The aim was to find all the measurements or control loops in a plant having the same behaviour because the root cause will be among that group. This section now examines methods available for diagnosis of the root cause.

Non-linearities such as valve friction constitute an important root cause of plant-wide disturbances (Bialkowski, 1992; Ender, 1993; Shinskey, 2000). Many vendors' publicity material emphasize these problems and Ruel of Top Control (Ruel, 2001) placed deadband, backlash and stiction at the top of the list of checks to be made prior to tuning a process control loop. Other

root causes also exist, however, including loop tuning faults and disturbances caused by interacting loops.

Normal operations versus special tests: The strong preference in the process industries is to conduct the first phase of detection and diagnosis using normal operating data without the need for any special tests. The concept was emphasized by Harris (1989) in the context of control loop performance assessment.

Special on-line tests may, however, be used for confirmation once a suspected root cause has been identified. Such tests initiate actuator movements (e.g. changes in valve position) by means of manual changes to the controller output (Åström, 1991; Gerry and Ruel, 2001). The control loop is disabled during the tests which means the flow through the valve is not maintained at an optimum value and therefore such special tests have the potential to disrupt production. Many control loops may be affected by a plant-wide disturbance and it is especially undesirable to apply the special tests one by one to them all in an attempt to find the one faulty valve (Ruel and Gerry, 1998). An aim of this thesis is to provide the initial diagnosis using normal operating data and thus to minimize the need for special tests by directing attention to the component most likely to be the root cause.

2.5.2 Diagnosis of individual loops

Op–mv plots: Many process control engineers use $x-y$ plots for the diagnosis of individual control loops. The controller output is plotted on the horizontal axis versus the manipulated variable on the vertical axis with time as parameter. The controller output is the demand to the valve and the manipulated variable is the flow through the valve or the valve position if this can be measured. The scaling on the axes is usually 0 to 100% for both variables and therefore if the control valve is healthy the plot is a straight line at 45°. Figure 7.1 and 7.2 in Chapter 7 show some examples of such plots for faulty valves with deadband and stiction where the plots assume various distinctive patterns. However, the valve positions are not normally available in the distributed control system and the wanted flows are often not measured at all which tends to limit the applicability of the $op-mv$ plots.

Horch (1999) showed further that the covariance between the controlled variable and the input to the controlled system (the manipulated variable) would be an odd function in the presence of a valve with deadband, a technique that works well for liquid flow loops when the controlled system is the valve. Horch reported that the relationship did not hold for control loops with integrating processes or long time constants such as level or pressure control loops.

Higher order spectral methods: Higher order spectra (Nikias and Petropulu, 1993) of the signals within a control loop can indicate the presence of a non-linearity (Emara-Shabaik *et. al.*, 1996;

Choudhury *et.al.*, 2002 and 2004; Choudhury, 2004). The Higher Order Spectra (HOS) toolbox of MATLAB provides the necessary routines.

A distinctive characteristic of a non-linear time series is the presence of phase coupling such that the phase of one frequency component is determined by the phases of others. Higher order spectra known as the bispectrum and bicoherence respond to quadratic phase coupling in a signal such as $x(t)$ below if the phase of the frequency component at $f_1 + f_2$ is equal to $\phi_1 + \phi_2$, but there is no bispectral response if ϕ_3 is a random phase:

$$x(t) = a_1 \cos(2\pi f_1 + \phi_1) + a_2 \cos(2\pi f_2 + \phi_2) + a_3 \cos(2\pi(f_1 + f_2) + \phi_3)$$

The squared bicoherence is a normalized quantity taking a value between 0 and 1:

$$bic^2(f_1, f_2) = \frac{|B(f_1, f_2)|^2}{E(|X(f_1)X(f_2)|^2)E(|X(f_1 + f_2)|^2)}$$

where $B(f_1, f_2)$ is the bispectrum at frequencies (f_1, f_2) given by:

$$B(f_1, f_2) = E(X(f_1)X(f_2)X^*(f_1 + f_2))$$

In the above expressions $X(f)$ is the Fourier transform of x at frequency f, $X^*(f)$ is its complex conjugate and E is the expectation operator.

The bispectrum has a non-zero value if there is significant phase coupling in the signal between frequency components at f_1 and f_2. The bicoherence gives the same information but is a value between 0 and 1 and is often presented as a surface plot similar to a spectrum but with two independent axes for f_1 and f_2. Peaks in the bicoherence at coordinates (f_1, f_2) indicate non-linearity. However, at the present time attempts to relate the shape and appearance of the peaks in the bicoherence plots to specific non-linearities such as stiction or deadband have not been successful.

SEVA technology: Self-validating (SEVA) instruments and actuators report a measurement value or a manipulated value and at the same time inform the controller and operator of the validity of the value (Clarke, 2000). The main group active in the area is the Invensys University Technology Centre at the University of Oxford and prototype SEVA instruments have been demonstrated such as a Coriolis flow meter, thermocouple and dissolved oxygen sensor (Yang and Clarke, 1997; Henry, 2001*a*). A SEVA web valve has also been demonstrated which reports it status through a web page so that maintenance staff can track the performance of control valves all over the plant (Henry and Clarke, 2004).

SEVA research is well funded and producing working prototypes but the industrial uptake has been slow. One reason is that they require Fieldbus wiring systems to be in place (Henry, 2001*b*) but these are not yet widespread. Henry has also identified that SEVA will require an open standard so that end users can mix and match products from different vendors, and the standard is not yet agreed.

The use of smart sensors and valves will become widespread in future because Invensys/Foxboro, Emerson Process Management and others are investing in these technologies. Once every instrument and valve has SEVA capability some of the diagnostic methods proposed in this thesis will be superseded because valves will report their own diagnosis. Detection of widespread disturbances and the isolation of root causes other than sticking valves would continue to be useful, however. Operating companies replace their control and automation systems infrequently however, and it is the author's opinion that all the methods being proposed in this thesis will continue to be needed for at least the next ten years.

2.5.3 Diagnosis of distributed disturbances

Root cause diagnosis: It is a challenging task to determine which single loop among many is the root cause of a distributed disturbance. The influences of a faulty control loop may propagate both downstream and upstream from the root cause because of recycles or the effects of control actions. For instance, the level in an upstream tank may be disturbed if the control valve on the (downstream) outflow causes a limit cycle oscillation. therefore plant layout alone is not a reliable guide.

Data-driven methods for non-linear root causes: Early studies used the presence of prominent harmonics as an indicator of non-linearity. The source of a limit cycle oscillation was distinguished from the secondary oscillations in a distillation column in a refinery (see Figure 1.1) and in a pulp and paper mill (Ruel and Gerry, 1998). It is not always true, however, that the time trend with the largest harmonics is the root cause. For instance, the second and third harmonics of a non-sinusoidal oscillatory disturbance are sometimes amplified in the secondary disturbance yet the linearity is still increased, as noted by Zang and Howell (2004). It can happen when a controller compensates for higher harmonics in a controlled variable (e.g. a level) by means of manipulation of a flow. The harmonics are then amplified in the flow.

Diagnosis of a non-linear root cause should be accomplished using an unambiguous signature for non-linearity that grows stronger closer to the source (Thornhill *et. al*, 2001). The scientific area relevant to this work is non-linear time series analysis (Kantz and Schreiber, 1997). The non-linearity test determines whether a time series could plausibly be the output of a linear system driven by Gaussian white noise, or whether its properties could only be explained as the output of non-linearity. The method is described in detail in Chapter 6.

Data-driven methods for other root causes: Other root causes of distributed disturbances include:

- Control loop interactions arising when two controllers have a shared mass and/or energy store (e.g. pressure and level controllers may compete for control of the contents of a reactor);

- Structural disturbances caused by coordinated transfers of mass and/or energy between different process units, especially when a recycle is present;
- Disturbances entering at plant boundaries (e.g. due to site utilities such as steam);
- Poorly tuned controllers.

The determination other types of root causes from normal operating data remains an open problem. To the author's knowledge, there exist no definitive signatures for the determination of root causes of plant-wide disturbances caused by control loop tuning, structural problems or loop interaction. Xia and Howell (2001*a*, 2001*b*, 2003) used a comparison of the signal-to-noise ratios of controlled variable and controller output to give a loop status that determined if an individual loop was affected by long or short term transients or a slow or fast oscillatory disturbance. Fast oscillatory disturbances were attributed to tuning or non-linearity.

Methods of advanced signal processing using causality analysis are starting to be used. The aim is to map out the structure of the process by using the measured data. An elementary case is when the time trend of Tag A correlates with a time-lagged trend of Tag B. It can be concluded that disturbances are propagating from A to B. Huang *et. al.*, (2002) used path analysis to determine direction as well as correlation in a data set. Path analysis (Johnson and Wichern, 1992) can distinguish between the case where A directly influences both B and C and the case where A influences B and B influences C.

Entropy measures are starting to be used to determine closeness and directionality. The term *entropy measure* implies that the analysis is being done using the probability density functions (*pdf*) of the measurements. Chiang and Braatz (2003) looked for changes in the *pdf* or changes in the relationships between the *pdf*'s of two measurements. They concluded that the information was most useful when combined with a knowledge of the process. Bauer *et. al.* (2004) applied transfer entropy calculations (Schreiber, 2000) to data from an industrial process and was able to detect the changes when a process fault occurred. The work explicitly addressed the problem of capturing time dependency with in an entropy framework. The ability to map out cause and effect paths in a plant from analysis of process data is potentially an attractive way forward for root cause diagnosis.

Qualitative model-based methods: Stanfelj *et al.* (1993) provided a decision-making tree which included cross-correlation between a feed forward signal and the controlled variable of the loop under analysis. Likewise, Owen *et al* (1996) showed an application of control-loop performance monitoring in paper manufacturing which accounted for upset conditions of the whole mill and interactions between control loops. These cases needed a knowledge of the process flowsheet, in particular about which loops might disturb one another. Chiang and Braatz (2003) made similar comments. The industrial case study in Chapter 8 of this thesis also required knowledge of the process flowsheet to complete the diagnosis. It is clear that the purely data-driven approach would be enhanced by the capture and integration of cause and effect information from a process schematic.

The trend in large petrochemical facilities, for instance in ethylene plants, is increasingly towards model-based control and optimization in which each plant has a first-principles process model based on the physics and chemistry of the process. The high cost of the model (see section 2.2.2) is justified for very large plants with narrow margins because of the potentially large profit from operating even fractionally closer to constraints. Models are also being more frequently used in polymer manufacture as vendors find accurate ways to adapt a basic model to different installations. Opportunities will therefore exist in future to exploit the linkage of process models and data-driven analysis for cause-and-effect reasoning.

2.6 CHAPTER SUMMARY

Chapter 2 has given an overview of the field of the thesis and placed it in the context of academic literature and the industrial state of the art. Three reviews by Venkatasubramanian *et. al.*, (2003*a*, *b*, *c*) highlighted several major areas and the work of the PhD was found to reside in the area that they termed *quantitative process history based methods.*

A review of vendor applications showed that in the area of chemical process control the process history based methods are more widely used than model-based methods for fault detection and diagnosis. The process sector contrasts with the situation in equipment, machinery and power systems where first principles models are used for diagnosis of machinery such as turbines and compressors. A further finding is that a distributed control system enables a user to collect process data in a systematic way and to control the plant. Applications making use of the data for fault detection and diagnosis tend to be supplied separately, often by specialist independent suppliers or consultants. The work of this thesis when developed commercially will be such an application.

A number of observations arose from the review to guide the work of the PhD programme:

- Spectral PCA was highlighted as solution for the problem of time delays and lags caused by process dynamics because the power spectra are invariant to time delays;
- The autocovariance function was found also to have a use for oscillation detection because it has the same period of oscillation as the time trend but is less noisy;
- Non-linearity is a common root causes of control problems, particularly non-linearity originating in valve faults such as a deadband due to friction. A non-linearity signature will be exploited to find a non-linear root cause of a plant-wide disturbance;
- Manual data preprocessing methods are depreciated because they reduce the industrial appeal of the techniques. Methods that are robust to gross errors are required;
- An automated means of detection of compressed data is needed because analysis of compressed data has been shown to give misleading results;
- There is scope in the future for linkage of the sections that Venkatasubramanian (2001) termed *process history based* and *qualitative model based.* The linkage has the potential to give enhanced root-cause reasoning.

3. DATA PRE-PROCESSING

3.1 OVERVIEW

3.1.1 Motivation

The previous chapter recommended that the amount of data pre-processing should be minimized to ease the transfer to industry of the methods of detection and diagnosis of distributed disturbances. Nevertheless, a preliminary inspection of data is invaluable and can help to guide the subsequent analyses in fruitful directions. The aim is to give a process control engineer a good overview of the main features present in the data. Items of interest are:

- Visual examination of time trends, power spectra and autocovariance functions of the plant measurements;
- Reporting of the means and standard deviations;
- Testing of the suitability of data for further analysis, particularly testing for the presence of compression;
- Filtering of selected frequency components, e.g. removal of high frequency noise or diurnal variations.

3.1.2 Introduction

Chapter 3 describes methods for visualizing plant data sets, checking them and preparing them so that they are suitable for use with the techniques for detection and diagnosis of distributed disturbances in the next three chapters.

The first subsection describes the essential step of viewing the data. A useful feature is to be able to observe numerous time trends at the same time for comparison with one another and a set of tools was created for performing this function. They create high density plots and enable a rapid visual inspection of key features of the time trends, spectra and autocovariance functions. Basic statistics of mean and standard deviation are also reported.

Section 3.3 describes a frequency domain filtering method which can be used for removal of low frequency non-steady trends, removal of high frequency noise, and close-up inspection of a plant-wide oscillation. A frequency domain filter is implemented based upon a specification of what components the user wants to retain in the signal. Real-time implementation is not required and

the filter can therefore exploit the whole data set , i.e. a non-causal filter can be used. The filter itself is one line of code in which the contents of unwanted frequency channels are set to zero.

Chapter 2 concluded that compressed data should not be used for data-driven analyses. A key data preparation step is to check the data for evidence of compression because the results would be meaningless if applied to compressed data. The justification of these statements, and an algorithm for compression detection will be presented in section 3.4.

3.1.3 Work done by the thesis author and others

Work reported in this chapter has been published in the following papers. This section describes the contributions made by the author and co-authors.

> Thornhill, N.F., Choudhury, M.A.A.S., and Shah, S.L., 2004, The impact of compression on data-driven process analysis, *Journal of Process Control*, **14**, 389-398
>
> Thornhill, N.F., Huang, B., and Zhang, H., 2003, Detection of multiple oscillations in control loops, *Journal of Process Control*, **13**, 91-100

Both papers were written by Thornhill while the co-authors read, discussed and criticized the text. Thornhill created all diagrams and graphs in the papers expect for Figure 6 in Thornhill, Choudhury & Shah (2003). The software was conceived and the results derived by Thornhill apart from the compression detection algorithm which was a joint effort with Shoukat Choudhury at the University of Alberta.

The concepts which are presented in this thesis originated with Thornhill. It is noted, however, that frequency domain filtering had independently been attempted by Huang and Zhang in order to remove high frequency noise from data trends.

3.2. DATA INSPECTION AND VISUALIZATION

3.2.1 Introduction

Aims: Visual examination and basic checking of the data is an important first step in any data analysis because the insights achieved from such examinations guide the selection of data analysis tools. This subsection outlines the methods for creating high density plots that aid visual

inspection and gives procedures for carrying out other basic checks on the data and for reporting their basic statistics.

High density plotting: High density plots display multiple time trends, autocovariance functions and power spectra on a single plot so that relationships between variables can be observed. Appendix A1 gives a brief account of the signal processing concepts and expressions needed for generation of the high density plots. Figure 3.1 shows an example of high density plotting of measurements from a plant at a SE Asian refinery. The chemical process is described in full at the start of Chapter 4 where several features of interest in disturbance detection and diagnosis are highlighted.

Basic data checking: Other basic checks on the data include removal of bad or missing values and calculation of basic statistics such as mean and standard deviation. If the measurements are from a control loop then it is useful also to know the rules that DCS vendors use to deal with the missing controller set points when control loops are being operated in Manual mode. These are explained in section 3.2.3.

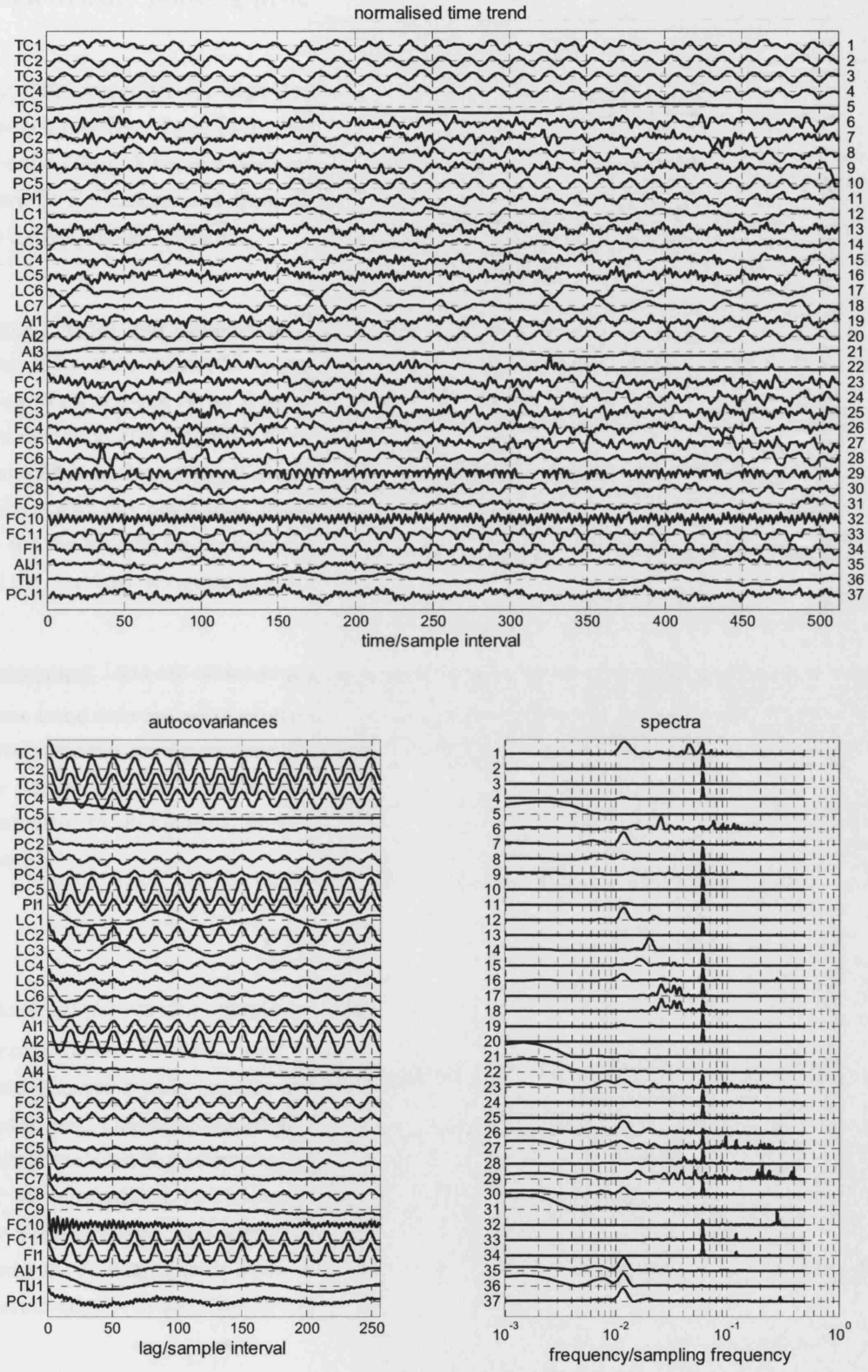

Fig 3.1 High density time trend plots, autocovariance functions and power spectra for measurements from a SE Asia refinery process described in section 4.2.

3.2.2 High density plotting procedures

Overview: The procedure for high-density plotting involves scaling of data and production of plots in which the plotted graphs are offset one below the other. The sources of data in a chemical plant are called *tags*. A tag is a label given by the distributed control system to one source of data which may be a measurement from an instrument (also called a process variable), a controller output or a manipulated variable.

Recording of means and standard deviations: It is important to record the means and standard deviations of the time trends. They are needed for many purposes, an important example is the monitoring of oil and gas flows from offshore oil platforms for tax purposes. Scaling of the time trends for detection and diagnosis of disturbances as specified below removes information about numerical values of the trends. It is impossible to tell from scaled time trend plots what were the mean values and the deviations from the mean. Therefore it is necessary to calculate basic statistics for the record, i.e. to determine the mean value, the standard deviation in engineering units and the standard deviation as a percentage of the mean value for each time trend.

Data preparation: Data are scaled to give a good visual impact, as follows:

- Time trend data are mean centred and scaled to unit variance;
- Power spectra are determined from mean centred data as the squares of the magnitudes of the discrete Fourier transform (DFT). Choices of scaling for good visual presentation are to normalize to unit power, or to scale so that the largest spectral peak is the same in all spectra:

$$\hat{P}(f_i) = \frac{P(f_i)}{\sum_{i=1}^{N} P(f_i)} \quad \text{or} \quad \hat{P}(f_i) = \frac{P(f_i)}{\max(P(f_i))}$$

  where $P(f_i)$ is the spectral power in the i'th channel and $\hat{P}(f_i)$ is its scaled representation. An issue with normalization to unit power is that spectra with features spread across a wide band of frequencies have magnitudes which are too small to be seen on the plot. Therefore largest-peak scaling is preferred for visualization;
- Autocovariance functions may be determined from the inverse discrete Fourier transform of the normalized two-sided power spectrum by application of the Wiener-Khinchin theorem, (Proakis and Manolakis, 1996). No scaling is needed if the autocovariance function is determined from the power spectrum scaled to unit power because the autocovariance at zero lags is then unity.

Plotting of time trends: Time trend data are stored as a vector of values $x(i)$, $i = 1$ to N. For instance, the vector may appear as a column in an Excel spreadsheet. In signal processing, the samples in a vector are counted from zero so for a time trend having N samples, the counter n

ranges from 0 to $N-1$. The position in the vector is $i=n+1$, thus $n=0$ refers to the first sample in the vector and $n=N-1$ is the N'th sample. The time tag t ranges from 0 to $(N-1)\times T_s$ where T_s is the sampling interval. The relationship between the sample counter n and the time tag is as shown in the top panel of Figure 3.2.

The horizontal axis in a high density time trend plot ranges from 0 to $(N-1)$ and is labelled *time/sample intervals*, consistent with the ISO31 Standard for declaration of numerical quantities. For example, an axis tick of 200 and an axis label saying *time/sample intervals* has the following interpretation:

$$time/sample\ intervals = 200 \quad \text{or} \quad time = 200\, sample\ intervals$$

If the sampling interval were, say, 2s it means the time at that position on the horizontal axis is 400s. The use of a sampling intervals axis is convenient for description of the period of oscillation or the duration of transient features (e.g. 250 samples per cycle).

For mean centred and scaled data an offset between trends of 5 units on the vertical axis is adequate to separate the trends so that each could be seen clearly. Thus the actual values plotted are $x_1(1:N)-5$, $x_2(1:N)-10$ ...., $x_m(1:N)-5m$ where $x_m(1:N)$ are the samples 1 to N of the m'th trend.

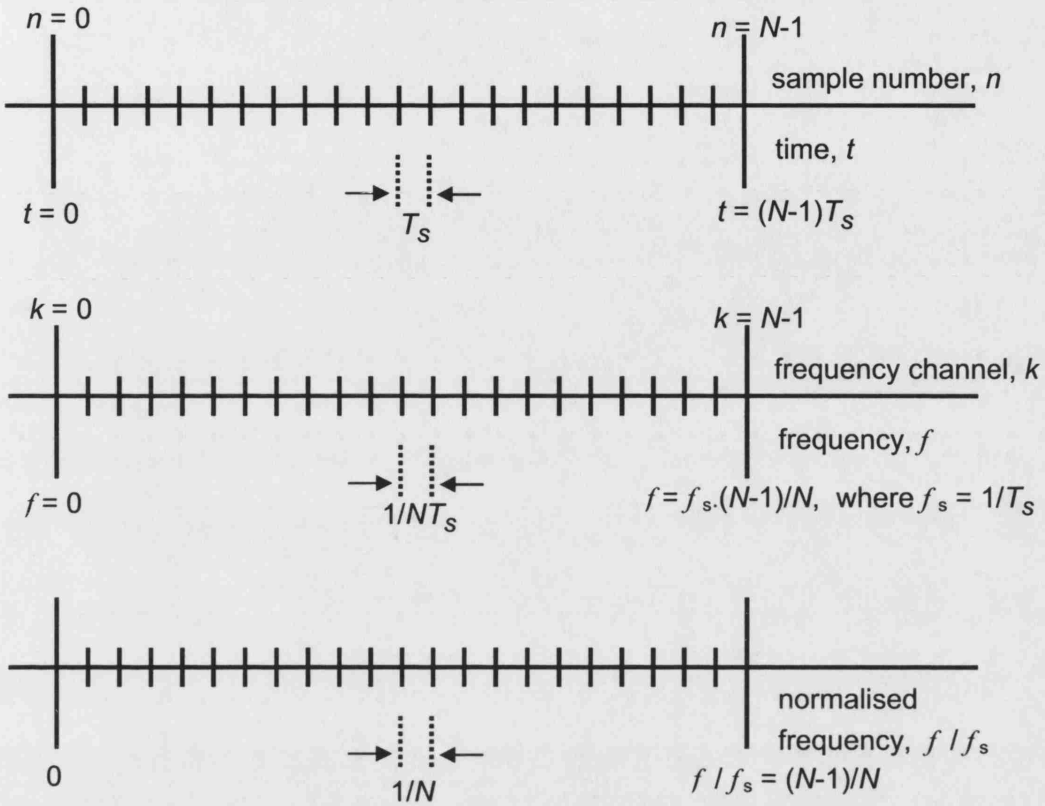

Fig 3.2 Upper plot: Time axis showing the relationship between sample number n and time t. Middle: Frequency axis showing the relationship between frequency channel k and frequency f. Bottom: A frequency axis expressed as a fraction of sampling frequency.

Plotting of power spectra: The power spectra (lower right hand panel of Figure 3.1) are plotted on a linear vertical axis and a logarithmic frequency axis. It is conventional to use the counter k for the frequencies, with the range of k being 0 to $N-1$ if the original time trend has N samples. The actual frequencies are given by $f_s \times k/N$, where f_s is the sampling frequency equal to $1/T_s$. It is common to refer to the frequencies by their $k-$values using the term *k'th frequency channel*. It is only valid to plot frequencies up to one half of the sampling frequency because of the symmetry and periodicity properties of the discrete Fourier transform caused by use of sampled data (see Appendix A1). This observation is equivalent to the Nyquist sampling theorem which requires at least two samples per cycle to characterize a sinusoidal signal. For even N, therefore, the spectral powers in frequency channels $k=0$ (the d.c. or zero frequency) up to $k=N/2$ are distinct and may be plotted. The spectral powers at frequencies $k=(N/2)+1$ to $N-1$ are the same as the powers for channels $k=N/2-1$ back down to $k=1$ and are not plotted.

The horizontal axis in the high density power spectrum plot is a fractional frequency axis and the top frequency plotted is 0.5. The axis label is $frequency/f_s$ where f_s is the sampling frequency

so that, for instance, the axis tick of 0.5 means $frequency = 0.5 \times f_s$. The benefit of fractional frequency scaling is that the frequency associated with a spectral feature can be easily determined because they are reciprocal. For instance, a peak at 0.01 has a frequency $0.01 \times f_s$ or 100 samples per cycle. It is then easy to cross reference such features in the spectra to oscillatory features having a period of 100 samples per cycle in the time trend plot with its *time/sample intervals* axis.

An offset of 1.2 units between spectra scaled to the same peak height is adequate to separate the spectra so that each can be seen clearly.

Plotting of autocovariance functions: The horizontal axis of the autocovariance functions is a linear lags axis. The counter for the lags is n and the actual lag is given by nT_s. In this work the autocovariance function is determined via the inverse Fourier transform of the power spectrum and therefore has the same periodicity properties as the power spectrum. For even N the autocovariance at lags $n = 0$ (no lag) up to $n = N/2$ are distinct and may be plotted but the autocovariances for $n = (N/2)+1$ to $N-1$ are a mirror image of the those for $n = N/2-1$ back down to $n = 1$ and are not plotted. An offset of 1.2 units between the autocovariance functions is adequate to separate the plots so that each can be seen clearly.

Comments on artifacts due to spectral leakage: The sample data sequence $\{x(n)\}$, $n = 0$ to N can be thought of as a subset of a sequence of infinite length multiplied by a rectangular window:

$$w(n) = \begin{cases} 0 & n < 0 \\ 1 & n = 0 \text{ to } N-1 \\ 0 & n > N-1 \end{cases}$$

The DFT of a windowed data sequence is the convolution of the DFT of the signal in the sampled data sequence and the DFT of the rectangular window, which is a *sinc* function having oscillatory side lobes (Mulgrew *et. al*, 2002; Proakis and Manolakis, 1996).

Spectral leakage occurs with a rectangular window unless the data sequence captures an exact number of full cycles of oscillation. Figure 3.3 shows the situation of a cosine wave where 7.5 cycles of oscillation were captured in 16 samples. The true power spectrum is a peak in channel 8 and its alias in channel 24, but the side lobes of the underlying *sinc* function (dashed line) cause the spectral power to spread across many frequency channels. The same situation applies in the high density power spectra reported earlier. Inspected closely, each spectral peak is broadened by the effects of spectral leakage unless the data sequence happens to contain en exact number of cycles of oscillation.

Other windows have less abrupt transitions and less pronounced oscillatory side lobes in the transformed domain and therefore may be preferred. For instance, the samples $\{x(n)\}$, $n = 0$ to N, may be multiplied by a Hanning window:

$$w(n) = 0.5 - 0.5\cos\left(2\pi\frac{n}{N}\right)$$

The Hanning window has a more gentle slope when plotted in the time domain and its transform $W(k)$ has reduced side lobes. Further window functions and their specifications and properties are given in Mulgrew *et. al* (2002) and reviewed in detail in appendix A1. Other windows generally reduce the effects of spectral leakage and therefore the reason for not applying windowing in the forward transform from the time domain to the frequency domain has to be defended. The reasons why the rectangular window is used are:

- For simplicity. The use of a rectangular window is automatic. By contrast other windows have to be explained to end users who then have to make choices which are a distraction from the work in hand;
- They are sufficient: Only large spectral features are of interest to process control engineers and these can be detected even in the presence of spectral leakage. If a signal is so small that it requires advanced spectral methods for its characterization then it is not affecting the economic performance of the plant. Even if detected, a process control engineer would not take action;
- The frequency domain filter of the next section uses a forward transform to create the spectra and inverse transforms to calculation filtered time trends and autocovariance functions. Any non-rectangular window introduces artifacts at the inverse transform stage.

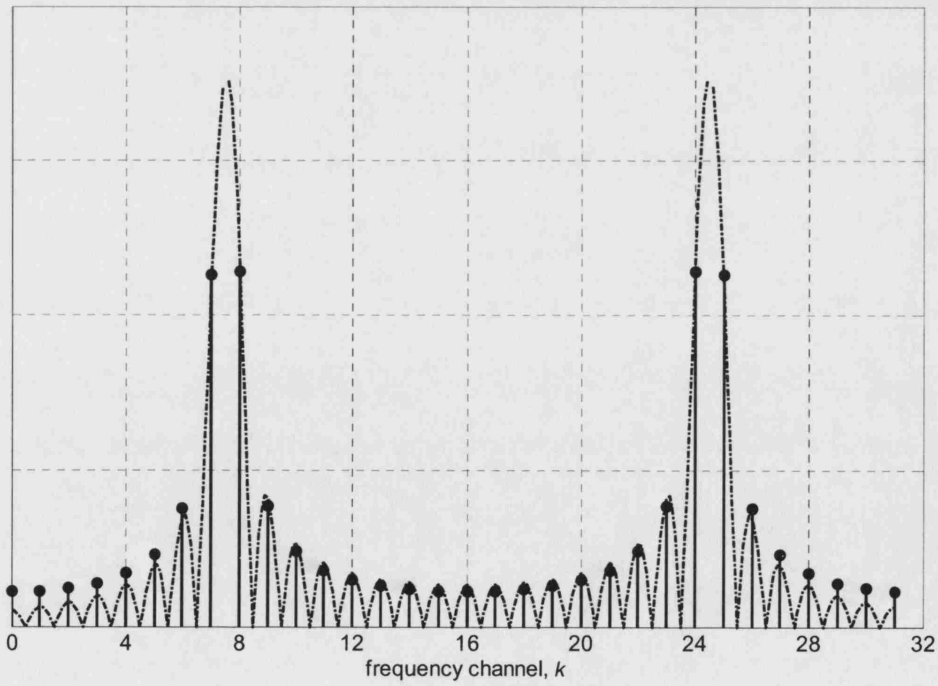

Fig 3.3 Two-sided power spectra with a rectangular window. The dashed line is the underlying *sinc* function, the black dots are the spectral power calculated from the DFT. The true spectrum should be a single peak in channel 8 and its alias in channel 24 but spectral leakage has destroyed that ideal pattern.

Inspection of the high density plots: Figure 3.1 showed a three panel high density plot of the time trends, autocovariances and power spectra of the process variables. This example has 37 tags. In other cases up to 50 tags have been displayed at a time, but then each panel has to be larger so that it occupies a full page rather than half a page. For showing more than 50 tags simultaneously it is recommended to use two or more separate plots.

Visual inspection of the high density plots show that a distributed oscillation is present in the plant. The time trends suggest there is a problem because several tags such as 2-4, 10, 11, 19, 20, 33 and 34 appear to have coordinated oscillatory patters of similar period. The benefit of the power spectra is that they give a sharper visual signature in the form of a spectral peak. A total of 22 tags show a distinctive sharp peak at about 0.06 on the normalized frequency axis (about 16 sample per cycle). These spectral peaks confirm the oscillation in tags 2-4, 10, 11, 19, 20, 33 and 34 and shows that other tags are also being upset by the same oscillation. A systematic way to discover the affected tags and the severity of the upset will be presented in Chapter 4.

The autocovariance plot illustrates why the autocovariance functions are useful. Tags that have time-domain oscillations, such as 2-4, 10, 11 and many others have oscillatory autocovariance functions with the same period of oscillation. The autocovariance function, however, is a much cleaner signal and the oscillations can be more easily confirmed visually. For instance, it is not clear from the time trends that tags 24 and 25 are upset by the oscillation with about 16 samples per cycle because those measurements are noisy. However, their *ACF* pots clearly show that the oscillation is present.

3.2.3 Other basic data checks

Bad values: Vendor data analysis tools generally incorporate data pre-processing, for instance the removal of invalid data. Bad data arise from data transmission errors and were signified by a code, for example a numerical value of -99999.99 would indicate an invalid value. As data transmission bandwidth and overall system reliably have improved the number of such instances the author has encountered in industrial data sets has reduced greatly in recent years and it is not now considered an issue requiring academic input. In the work in this thesis bad values were set to NaN (not a number) which enables them to be processed in a consistent and predictable manner by any numerical program that conforms to ANSI/IEEE Standard 754-1985, (Standard for Binary Floating Point Arithmetic).

Basic statistics: The basic statistical calculations report the standard deviation to two significant figures and the number of significant figures retained for the mean is adapted to match the uncertainly indicated by the standard deviation. The result of this procedure is that measurements whose standard deviations are a small percentage of the mean value are reported to more

significant figures. Thus a raw result of 14195.43 ± 1487.42 is reported as 14200 ± 1500 while a raw result of 18467.3 ± 20.566 becomes 18467 ± 21.

Permission for use of the SE Asian data has only been given to show the mean centred values scaled to unit standard deviation, therefore it is not possible to show the table of results for the basic statistics.

Loops in manual and cascade loops: Data sets in a DCS or data historian include a loop status tag. The main classifications are AUTO (loops in automatic mode, the normal status), MAN (loops in manual under direct operator control) and CAS (cascade loops receiving their set points from another control loop).

If the loop status is MAN or CAS then the DCS vendors apply the following rules:

- If the loop status is MAN then the operator not the controller is setting the loop *op* signal. The *op* signal is normally constant unless the operator is doing a startup or some special tests;
- If the loop status is MAN then the loop set point tag (*sp*) is set equal to the process variable (*pv*) by the DCS where the *pv* is the measured value of flow, temperature and so on;
- If the loop status is CAS then the set point (*sp*) of the slave loop is equal to the controller output (*op*) of the master loop.

If the loop status is not provided with a data set then it can be deduced by reverse application of the above rules. For instance, a loop whose *sp* and *pv* are identically equal to one another is being operated manually.

An anomalous situation is sometimes observed in which one loop appears to be in AUTO but its controller output (*op*) is identically equal to the *pv* of another loop. The reason for this observation is that the loop is a master loop in a cascade where the slave loop is in MAN mode. Since the slave loop status is MAN its *sp* is set by the DCS to be equal to its *pv*. Since the master loop *op* has to be equal to the *sp* of the slave loop the DCS then sets the *op* of the master loop to the *sp* of the slave loop. That is how the *op* of one loop can appear to be equal to the *pv* of another loop.

3.3 FREQUENCY DOMAIN FILTERING

3.3.1 Introduction

Aims: The purpose of filtering of data is to remove noise or interference that might mask features of interest in the spectra, time trends and autocovariance functions. It is convenient for a user to specify the filter boundaries in terms of samples per cycle because that quantity can be easily seen from the time trend plot. For example, a filter is applied shortly to retain spectral features having 10 to 30 samples per cycle (spectral features from 0.033 to 0.1 on the normalized frequency axis) in order to focus on oscillations with about 16 samples per cycle (0.06 on the normalized frequency axis). All other frequencies are removed from the signal by this filter. With such filtering it is possible to determine whether the 16 samples per cycle oscillation is present even when it is dominated by a much larger amplitude oscillation at a different frequency, or when it is swamped by noise.

Non-causal filtering: The focus of this thesis is plant auditing using historical data. Therefore it is not necessary to use a real-time or on-line filter and the advantages of non-causal filtering can be exploited. In general, non-causal filters have the following benefits:

- They provide better smoothed estimates because all the N data points in the time trend $x(n)$ can be used to determine the filtered values. In particular, $x(n), n = m+1 \ldots\ldots N$ as well as $x(n), n = 1 \ldots\ldots m$ can be used to determine a filtered value $\tilde{x}(m)$. This is not possible in an on-line filter if m is the present time because $x(n), n = m+1 \ldots\ldots N$ would be in the future.
- The filter can be designed to have a zero phase characteristic and thus to preserve the phase relationships between frequency components in the filtered signal. This is important for the diagnostic purposes where it can be important to retain the shape of a non-sinusoidal oscillation. For instance, a square-ish waveform is characteristic of flow control loops with a sticking valve, but changing the phases of the harmonics of a square wave would destroy the shape of the square wave even if the amplitudes do not change.

3.3.2 Filter formulation

Removal of unwanted frequencies: The first step in frequency domain filtering is to calculate the discrete Fourier transforms (DFT) of the time trends of the process measurements. The frequency-domain filter then sets the values of the discrete Fourier transform $X(k)$ to zero in unwanted frequency channels.

A property of the DFT is that $X(N-k)=X^*(k)$, where * represents the complex conjugate. Therefore the magnitude of the DFT in channel $N-k$ is the same as that for channel k, where N is the number of frequency channels and k is the counter taking values 0 to $N-1$ (N is even in this formulation). The frequency domain filter must operate on the required channels and also on the aliased channels. Suppose $N=16$ and the filter is a high pass filter designed to retain 4 samples per cycle up to 2 samples per cycle (0.25 to 0.5 on the frequency axis). The wanted channels are $k=4$ to $k=7$ together with the Nyquist frequency ($k=8$) and the aliased channels $k=9$ to $k=12$. The filter is thus

$$\Phi(k)=\begin{cases}0 & k=0-3\\ 1 & k=4-12\\ 0 & k=13-15\end{cases}$$

and the filtered DFT is:

$$\tilde{X}(k)=X(k)\Phi(k)$$

The normalized filtered power spectrum is:

$$\tilde{P}(k)=\frac{\left|\tilde{X}(k)\right|^2}{\sum_{k=0}^{N-1}\left|\tilde{X}(k)\right|^2}$$

Filtered time trends $\tilde{x}(n)$ ($n=0$ to $N-1$), are reconstructed from the inverse DFT of $\tilde{X}$:

$$\tilde{x}(n)=\frac{1}{N}\sum_{k=0}^{N-1}\tilde{X}(k)e^{j2\pi nk/N}$$

while the filtered autocovariance function is the inverse DFT of the normalized filtered two-sided power spectrum and is scaled so that $\widetilde{ACF}(0)=1$:

$$\widetilde{ACF}(n)=\sum_{k=0}^{N-1}\tilde{P}(k)e^{j2\pi nk/N}$$

The filter is a simple and approximate realization of a non-causal Wiener filter (Press, *et. al.*, 1986; Proakis and Manolakis, 1996). The simplification involved is that a true Wiener filter also requires an estimate of the noise power within the wanted frequency channels which would then be subtracted from those channels.

Optimized plotting of the ACF: For enhanced visualization of the autocovariance functions the range of the axis is adjusted to match the filtering applied. For instance, if features having between 20 samples per cycle and 100 samples per cycle are present in the filtered autocovariance function then the maximum x-axis value would be set to $4\times100=400$ lags.

3.3.3 Design considerations

Artefacts in frequency domain filtering: Spurious results arise if the filter width is too narrow. For instance, a filter with a width of just one frequency channel would give a pure sinusoidal output even if the input signal were random noise. The filter $\Phi(k)$ has a rectangular shape and therefore becomes an oscillatory *sinc* function after the inverse transformation to the time domain or autocovariance function. Guidelines for the minimum filter width are now given such that the filtered time domain and autocovariance functions will not show spurious oscillatory behaviour.

The derivation is done using continuous rather than sampled data because the manipulation of trigonometric functions is more straightforward, the continuous result is then transformed into an equivalent recommendation for the sampled data case.

Minimum filter width Δf: Consider a flat power spectrum of unit magnitude $P(f)=1$ filtered by a two-sided filter of width Δf centred at $\pm f_o$. This section shows that the condition for the minimum width of the filter to avoid oscillatory artefacts is:

$$\frac{2}{\Delta f}\leq\frac{5}{f_o} \quad \text{or} \quad \Delta f\geq\frac{f_o}{2.5}$$

The filter has a rectangular shape, having a value 1 in the pass band and zero at all other frequencies. The *ACF* at lag τ is the inverse Fourier transform of the filtered two-sided power spectrum $P(f)$:

$$
\begin{aligned}
ACF(\tau) &= \int_{-\infty}^{\infty} P(f) e^{2\pi j f \tau} df \\
&= \left(\int_{-f_o-\Delta f/2}^{-f_o+\Delta f/2} e^{2\pi j f \tau} df + \int_{f_o-\Delta f/2}^{f_o+\Delta f/2} e^{2\pi j f \tau} df \right) \\
&= \frac{1}{\pi\tau 2j} \left(\begin{array}{l} e^{-2\pi j f_o \tau} \left(e^{2\pi j \frac{\Delta f}{2}\tau} - e^{-2\pi j \frac{\Delta f}{2}\tau} \right) \\ \qquad + e^{2\pi j f_o \tau} \left(e^{2\pi j \frac{\Delta f}{2}\tau} - e^{-2\pi j \frac{\Delta f}{2}\tau} \right) \end{array} \right) \\
&= 2 \times \cos(2\pi f_o \tau) \times \frac{\sin(2\pi\tau \Delta f/2)}{\pi\tau}
\end{aligned}
$$

If the filter is narrow then Δf is small and the autocovariance function becomes $ACF(\tau) \approx 2\Delta f \cos(2\pi f_o \tau)$. This ACF is an oscillatory function of τ even though the power spectrum before filtering was flat with no spectral peaks. Therefore it is undesirable that Δf be small. For large values of Δf the $\sin(2\pi\tau\Delta f/2)$ term disrupts the regularity of the $\cos(2\pi f_o \tau)$ term and the spurious oscillations therefore disappear.

The $\cos(2\pi f_o \tau)$ term has zero crossings when $\tau = (2m-1)/4f_o$ where m is a positive integer. The interval between the first and eleventh zero crossing is $5/f_o$. Eleven is the key number because in Chapter 4 an ACF is deemed to be oscillatory if the ten intervals between the first eleven zero crossings are regular. However, it is required that the ACF of the filtered flat spectrum *not* be oscillatory. Therefore Δf must be large enough that the zero crossings of the $\sin(2\pi\tau\Delta f/2)$ term disrupt the regularity of the first eleven zero crossings of the $\cos(2\pi f_o \tau)$ term. The $\sin(2\pi\tau\Delta f/2)$ term has zero crossings when $\tau = m/\Delta f$. Let the requirement be for at least two zero crossings of the $\sin(2\pi\tau\Delta f/2)$ function before $\tau = 5/f_o$. Therefore the condition for the width of the filter is:

$$\frac{2}{\Delta f} \le \frac{5}{f_o} \quad \text{or} \quad \Delta f \ge \frac{f_o}{2.5}$$

<u>*Minimum width in a sampled system*</u>: In a sampled data system the filter would have the following form:

$$
\Phi(k) = \begin{cases} 0 & k = 0 \text{ to } k_1 - 1 \\ 1 & k = k_1 \text{ to } k_2 \\ 0 & k = k_2 + 1 \text{ to } N - k_2 - 1 \\ 1 & k = N - k_2 \text{ to } N - k_1 \\ 0 & k = N - k_1 + 1 \text{ to } N - 1 \end{cases}
$$

where $k_1 < N/2$ and $k_2 \leq N/2$. Its width is $k_2 - k_1$ and the centre is $k_o = \left(k_1 + k_2\right)/2$. The relationship with the frequencies involved in the previous continuous analysis are:

$$\Delta f = \left(k_2 - k_1\right) \times \frac{1}{NT_s}$$

and

$$f_o = k_o \times \frac{1}{NT_s}$$

where T_s is the sampling interval and N is the number of samples in the original $x(n)$ sequence in the time domain. The specification for the width of the filter in terms of frequency channels is therefore:

$$\left(k_1 - k_2\right) \geq \frac{k_o}{2.5}$$

For instance, if $N = 128$ then $k = 64$ is the Nyquist frequency corresponding to $0.5 f_s$ or 2 samples per cycle. A filter with $k_2 = 32$ and $k_1 = 16$ would retain frequencies in the range $0.125 f_s - 0.25 f_s$, i.e. oscillatory features having between 8 samples per cycle and 4 samples per cycle. For that filter, $k_2 - k_1 = 16$ and $k_o = \left(k_1 + k_2\right)/2 = 24$. Therefore the filter is wide enough to avoid the spurious oscillations because 16 is greater than $24/2.5 = 9.6$.

Assessment of the strength of an oscillation: The power spectrum is normalized before filtering, to give $\sum_{k=0}^{N-1} \hat{P}(k) = 1$. Thus, when the spectrum is filtered it follows that the summed spectral power in the selected frequency channels gives the power as a fraction of the total. These power values include both oscillatory components and noise because the simplified Wiener filter does not subtract noise from the wanted frequency band. However, if there is a regular oscillation within the filter band the summed power in the filtered spectrum would be dominated by the power of the oscillatory signal and may be utilized as an approximation to the power of the oscillation.

Non-rectangular window filter functions: The artefacts resulting from the use of the rectangular shape of the frequency domain filter could be ameliorated by use of a non-rectangular window. The rectangular frequency domain filter was:

$$\Phi(k) = \begin{cases} 0 & k = 0 \text{ to } k_1 - 1 \\ 1 & k = k_1 \text{ to } k_2 \\ 0 & k = k_2 + 1 \text{ to } N - k_2 - 1 \\ 1 & k = N - k_2 \text{ to } N - k_1 \\ 0 & k = N - k_1 + 1 \text{ to } N - 1 \end{cases}$$

It has abrupt transitions from 0 to 1 and from 1 to 0. As discussed earlier, other windows have less abrupt transitions and less pronounced oscillatory side lobes in the transformed domain. Non-rectangular windows have drawbacks however. These are:

- Lack of clarity: Non-rectangular windows have to be explained to end users who then have to make choices which are a distraction from the work in hand. In particular the specification of a non-rectangular window in terms of the frequencies being cut out would need careful thought.
- A non-rectangular window has a processing loss because the area of the power spectrum under a non-rectangular window is not the same as the area of the rectangular window. Therefore non-rectangular widows cannot be used to determine percentage power in a frequency range.

In the author's opinion the drawbacks outweigh the advantages and therefore the rectangular frequency domain filter is recommended. It may not be optimal but it fits the purpose.

3.3.4 Demonstration and worked example

Choosing the filter: This subsection gives a demonstration of frequency domain filtering. The same data from the SE Asian refinery are used. Figure 3.1 showed evidence for a plant-wide oscillation with a period of about 16 samples per cycle (about $0.06f_s$ on the frequency axis). The aim of filtering is to remove the noise interferences at lower and higher frequencies in order to find the strength of the oscillations in the various tags and to observe the time trends and autocovariance functions of the wanted components. The filter applied retains spectral features with 10 to 30 samples per cycle ($0.033f_s$ to $0.1f_s$). The spectral power at all other frequencies has been set to zero, as can be seen in the plots of the filtered spectra in Figure 6.4. All the user has to do is to type [10 30] to specify the filter.

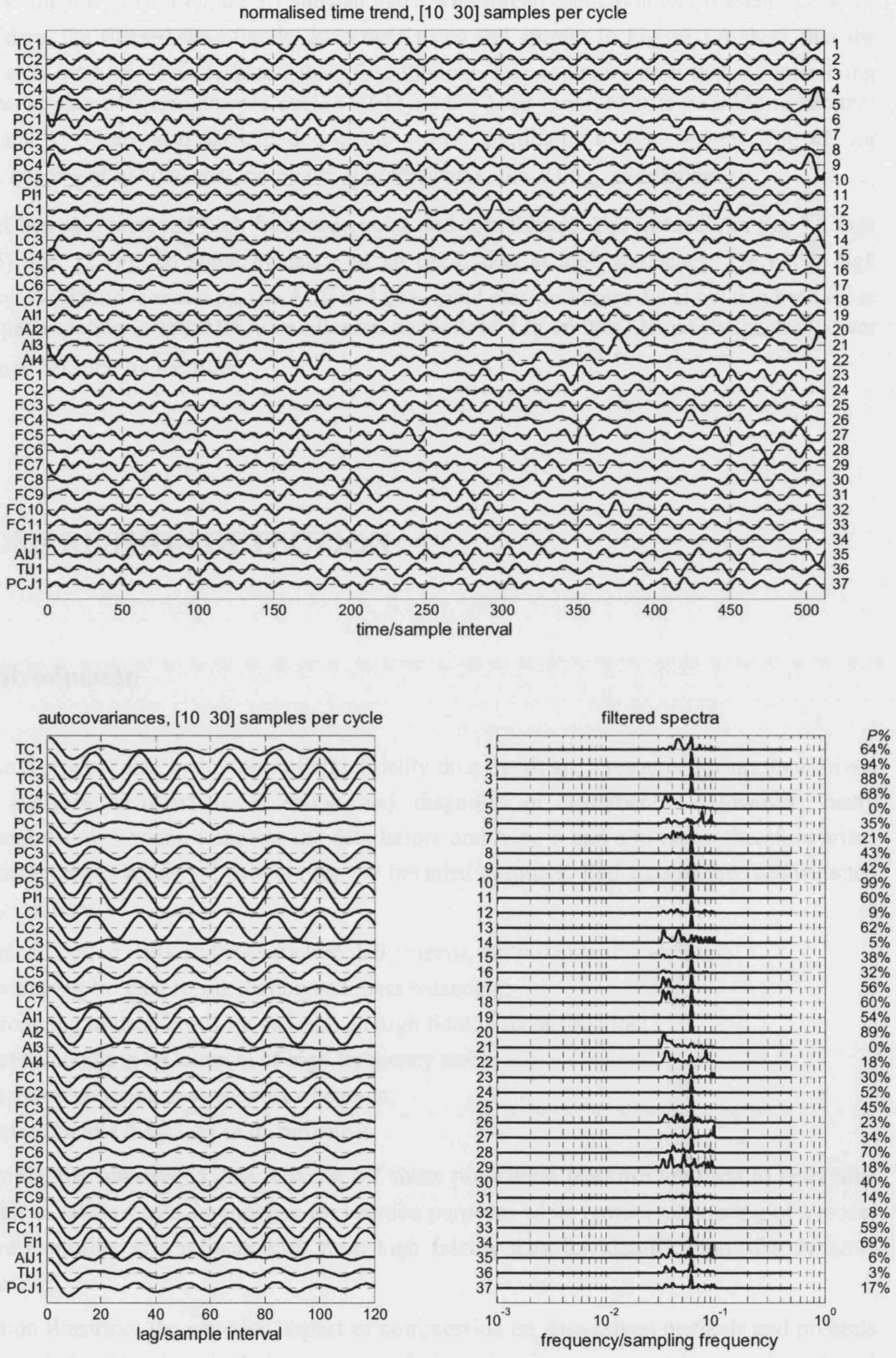

Fig 3.4 High density time trend plots, autocovariance functions and power spectra for measurements from filtered SE Asia refinery data.

Results: While the plots in Figure 3.1 showed that a widespread oscillation was present in this set of plant data, the filtered time trends, autocovariances and spectra in Figure 3.4 show that the oscillation is even more widespread than first thought. For example, now that an interfering oscillation at about 90 samples per cycle ($0.011f_s$) has been removed it is possible to see that tags 12 and 35-37 are also participating in the $0.06f_s$ oscillation to some extent. Tag 12, for instance, has 9% of its total spectral power concentrated in the $0.06f_s$ oscillation.

The filter has also removed high frequency noise and interference. For instance, in Fig 3.1 tags 16 (LC5) and 32 (FC10) show evidence of an oscillation at high frequency. Once the high frequency oscillation is removed, the ACF and time trend plots in Figure 3.4 show that both those tags are participating in the $0.06f_s$ oscillation, and indeed Tag 16 has 32% of its spectral power in the range selected by the filter.

3.4 COMPRESSION DETECTION

3.4.1 Introduction

Aim: Stored process data in the form of high fidelity time trends are a resource for the data-driven process analyses involved in detection and diagnosis of disturbances. However, many commercial data historians compress the data before archiving it and a question therefore arises of how useful the compressed data are for the intended purposes. End uses of the reconstructed data may include:

- Calculation of daily statistics such as daily means, daily standard deviations;
- Averaging for data reconciliation and mass balancing;
- Archiving of data trends for subsequent high fidelity reconstruction;
- Data smoothing by removal of high frequency noise;
- Feature extraction and recovery of events;
- Detection and diagnosis of disturbances.

For example, the transmitted data from an off shore production platform are used to determine daily totals of oil flow into the pipeline for taxation purposes while remote monitoring of a model predictive controller at a refinery may need high fidelity data for identification of a dynamic process model.

This section illustrates the negative impact of compression on data-driven methods and presents an automated algorithm by which the presence of piecewise linear compression may be inferred during the pre-processing phase of a data-driven analysis. The study shows that compression

interferes with many types of data-driven analyses and therefore an essential pre-processing step is to check whether compression is present.

It is noted that the information in certain types of process trends may be unaffected by compression. Constant values or infrequently changing values such as set points, targets and high and low limits could all be accurately reconstructed from compressed data.

Motivating compression example: Compression using piecewise linear trending is in widespread use in industrial data historians. For instance, AspenTech described an adaptive method based upon the box-car/backward slope method (AspenTech, 2001) while OSI state that their PI data historian uses a type of swinging door compression algorithm involving a compression deviation blanket with a width equal to twice the compression deviation specification (OSI Software Inc, 2002).

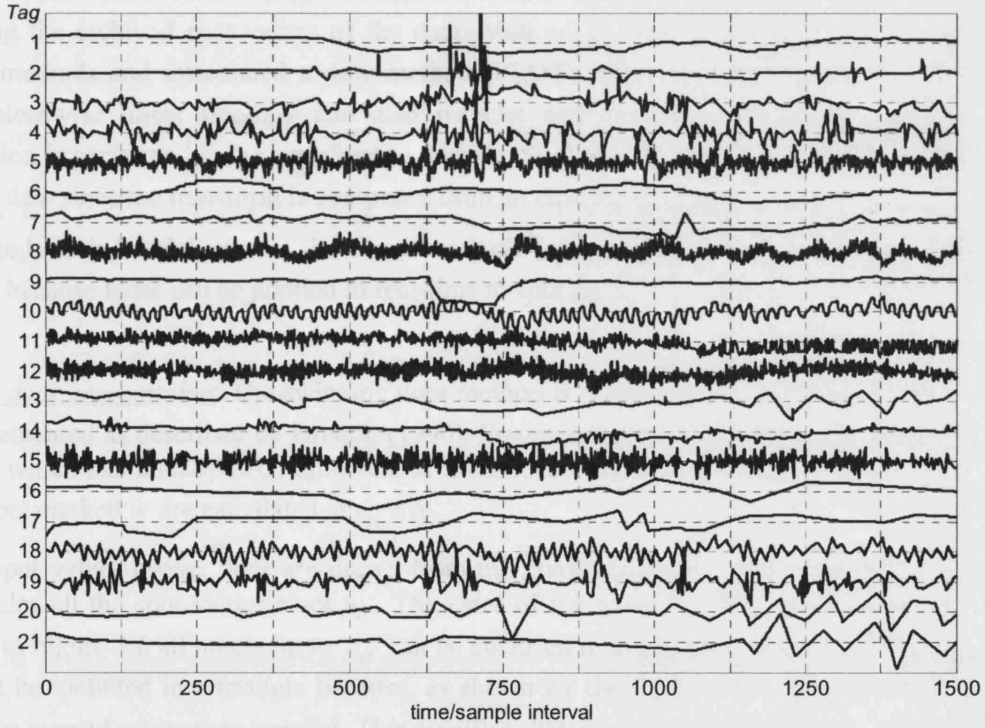

Fig 3.5 An industrial data set with compression in some tags, courtesy of Celanese Canada. Time trends are mean centred and normalized.

Figure 3.5 shows a data set from a data historian typical of those from which engineers might wish to extract useful information (courtesy of Celanese Canada Inc.). The straight line segments characteristic of industrial data compression can be seen in many of the time trends. It will be shown that compression factors of up to 94 were in use. The original uncompressed data were lost forever when the they were compressed and archived and it is now not possible to determine

what features have been lost. Later sections will show that most of these data trends are too compressed and that data-driven process analysis would, if attempted, give a misleading indication of the results that the original data would have given.

3.4.2 Compression methods

Overview of methods: Compression techniques can be divided into two main functional groups, direct methods and transform methods. A direct method makes the archiving decision in real time as the data are captured from the process. The box-car/backward slope (BCBS) method (Hale and Sellars, 1981) and the swinging door method (Bristol, 1990) are direct methods that use heuristic rules to decide whether to archive a spot value and the rules are tuned to achieve the capture of exceptions and linear trends. They reconstruct a data trend as a series of linear segments connecting the archived spot values of the data. Mah *et. al* (1995) compared piecewise linear trending methods and introduced a new method (PLOT). The work of Watson *et. al.* (1998) studied piecewise linear trending and also wavelet and Fourier compression in which the compression is performed in the transformed domain. Such methods are not real-time and require historical data since the transform is computed from an ensemble of data.

The methods in industrial use are the direct methods (also known as piecewise linear trending methods) because these can be applied in real-time to spot data.

Swinging door compression: The swinging door method is representative of industrial practice. It was implemented as described by Bristol, (1990). Figure 3.6 shows the principle. The first black point y_a, which has already been archived, is taken to be the start of a trend. Upper and lower pivot points marked × are calculated at $y_a \pm \Delta$.

As new spot values arrive, lines are drawn from the pivot points to form a triangular envelope that includes all the spot values since y_a. The sides of the triangle are the swinging doors. For instance, in Figure 3.6 all points up to y_d can be enclosed in a triangle. However, the next point, y_e cannot be included in a triangle because, as shown by the dashed line, the upper and lower doors have opened wider than parallel. This signifies that a new trend started at y_d. Point y_d is archived and the procedure starts again at y_d.

The trends are reconstructed from the archived spot values by linear interpolation between archived points at the original sampling instants. The compression factor (CF) for swinging door compression is defined as the ratio between the storage requirement of the original data set and that of the archived data. If the original data set had 1000 observations and 1000 time tags a direct method with $CF = 10$ would yield 100 observations, 100 time tags and 99 linear segments.

The compression factor is not specified directly in swinging door compression. Instead, the parameter to be set is the deviation threshold Δ in engineering units. Therefore in conducting the

compression tests described in this paper it was necessary to first conduct calibration trials to find the deviation thresholds corresponding to $CF = 10$ for each data set.

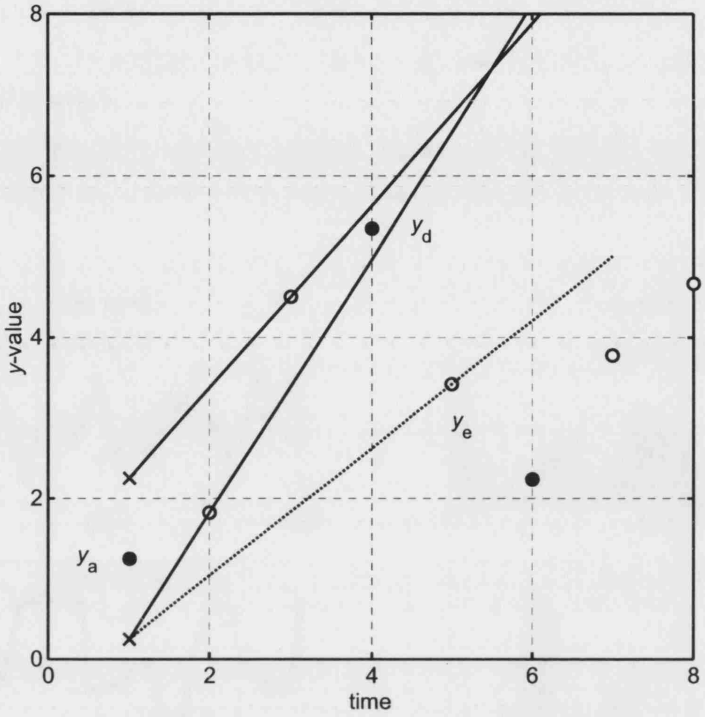

Fig 3.6 Swinging Door compression. Black circles represent archived spot values, values with open circles are not archived. At time step 5 (point y_e) the lower door (dotted line) opens up wider than parallel showing that a new trend started at y_d.

3.4.3 Examination of compressed data

Description of data sets: The effects of compression were studied on three uncompressed time trends of refinery liquid flow measurements, courtesy of BP Oil. The reason for selecting these particular examples is that they have contrasting characteristics. Each data set comprised nearly three hours of 10s samples representing deviations of flow in a process stream from the mean value. The left hand panels of Figure 3.7 shows portions of the time trends (the open circles) while the black bars in the right hand panels show their power spectra on linear vertical and horizontal axes. Numerical scales on the y-axis were omitted by request of the company.

Data set 1 has a persistent oscillation characterized on average by about twenty samples per cycle and its spectrum has a broad peak at a frequency of 0.05 times the sampling frequency (i.e. 20 samples per cycle). The challenge for high fidelity compression and reconstruction is to retain the spectral peak in the frequency domain and the oscillatory features in the time domain.

Data set 2 has a tendency to stay at a value for a time and then to move rapidly to a new level. It is from a control loop which has a limit cycle caused by a sticking valve. The signal is predictable for long periods and its spectrum shows very low frequency features because the period of oscillation is long while a series of harmonics highlights the non-sinusoidal nature of the waveform. The low frequency features and harmonics should be preserved during compression and reconstruction.

Data set 3 has little predictability and has spectral features at all frequencies. This signal is dominated by random noise and is from a well tuned loop operating close to minimum variance.

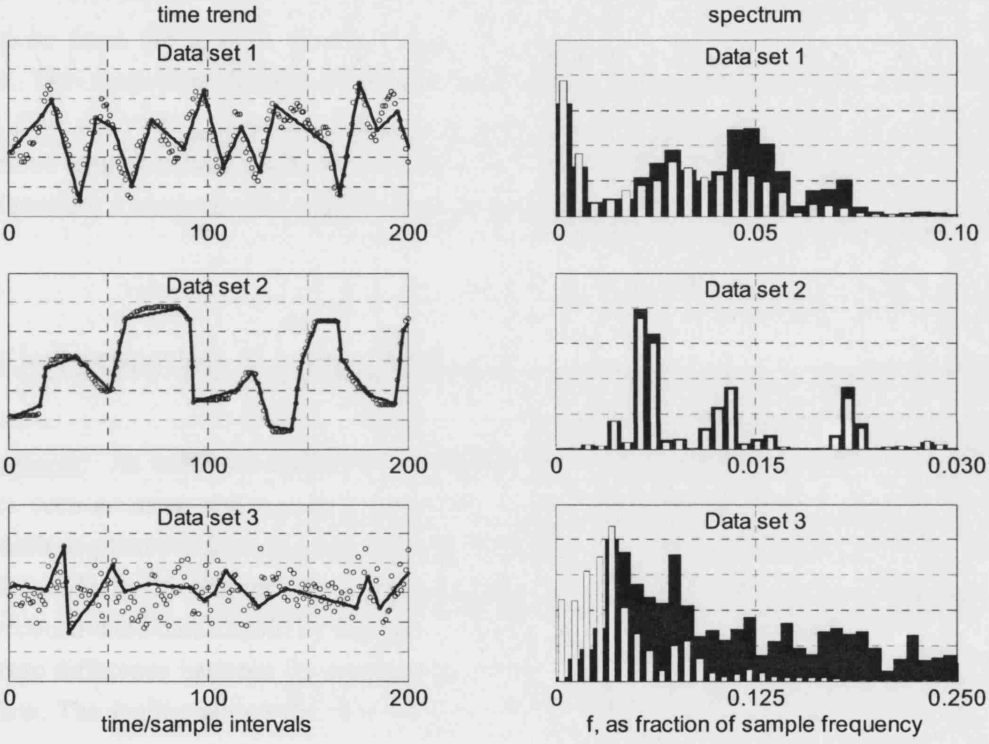

Fig 3.7 Left: Time trends and reconstruction with compression factor 10. Right: Original spectra (black) and spectra of reconstructed signals (white).

The effects of compression: Figure 3.7 shows the data(open circles) and reconstructions (solid line) with compression factor of 10 for data sets 1 to 3. Each complete data set had 1024 samples. Features of note are:

- The reconstructed trends do not follow all the oscillations in data set 1 because with $CF = 10$ the average duration of each linear segment is longer than half of the oscillation period;
- High fidelity compression was possible with data set 2 but with data set 3 much of the randomness is lost from the reconstructed trends.

The right hand panel in Figure 3.6 shows reconstruction in the frequency domain. The power spectra of the original signal are in black with the spectra of the reconstructed signals with $CF = 10$ overlaid in white. When the two are not the same then there is a reconstruction error.

- The spectral feature in data set 1 at 0.05 samples per cycle was not fully captured by the reconstructed data set;
- Data sets 1 and 3 had errors at low frequency and a non-zero spectral error at $f = 0$. Therefore the signals reconstructed after compression have different mean values than the originals;
- The low frequency harmonics of data set 2 are reproduced well but the high frequencies of data set 3 were not captured.

The observations from the spectra reinforce and illuminate the observations from the time domain plots. The frequency domain plots also give insight into why data set 2 is more compressible than data set 3. Data set 2 has very few spectral features and they are at low frequency (i.e. of long duration) while data set 3 has features over the whole frequency range. Data set 2 is therefore a much simpler signal with fewer different types of behaviour to capture.

3.4.4 Statistical properties of compressed data

Statistical measures: As indicated earlier, it is necessary in a data-driven analysis to calculated basic statistics such as mean and standard deviation. These statistics have uses in steady-state assessments such as plant production rates, data reconciliation and mass balancing, for instance for the detection of leaks. If compressed archived data are to be used for these purposes the mean value of the reconstructed data should be the same as the mean of the original. The measure used is the percentage difference between the mean values (PDM) scaled by the standard deviation of the original data. The scaling allows the relative significance of any change in mean value to be assessed:

$$PDM = 100\frac{mean(y) - mean(\hat{y})}{\sigma_y}$$

Two variance measures are also examined. The measures used are the ratios between the variance of the reconstructed data ($\sigma^2_{\hat{y}}$) and the variance of the original data (σ^2_y) (*RVC*), and between σ^2_y and the variance of the reconstruction error σ^2_e where $e_i = y_i - \hat{y}_i$ (*RVE*). The measures are:

$$RVC = \sigma^2_{\hat{y}} / \sigma^2_y \qquad \text{and} \qquad RVE = \sigma^2_e / \sigma^2_y$$

If the *RVC* and *RVE* measures sum to 1 then the reconstruction error is the orthogonal complement of the compressed signal (i.e. the sequence $y_i - \hat{y}_i$ is uncorrelated with the sequence $\hat{y}_i$). The significance of this observation is considered shortly.

Effects of compression on statistics: Figure 3.8 shows the behaviour of the mean value and variance measures as a function of compression factor. Noteworthy observations are:

- The mean of the signal reconstructed from the compressed archive differs from the mean of the original;
- The variances of the reconstructed data are smaller than the variance of the original signal;
- The sum of *RVC* and *RVE* at a given compression factor is not equal to 1.

These results show that data compression gives misleading information about basic statistical properties of the data. Compression alters both the mean and variance. Although the changes in the means are only a small percentage of the standard deviation they can have a big impact. For instance, the purpose of data reconciliation is often to find small shifts in the mean value that may be indicative of problems such as leaks. The shift in mean due to data compression may therefore be wrongly interpreted as evidence of a leak. Decisions of the type used in statistical process control will also be wrong if the warning and alarm limits have been based upon a statistical distribution determined from compressed archived data.

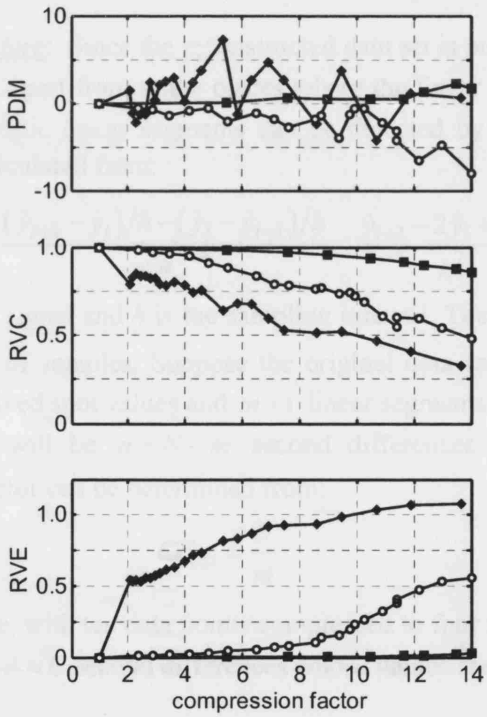

Fig 3.8 Statistical measures as a function of CF for data set 1 (circles), data set 2 (squares) and data set 3 (diamonds).

The sum of the measures of error variance (*RVE*) and compressed signal variance (*RVC*) is not 1. Thus there exists a correlation between the part of the signal deleted during compression and the compressed signal itself. The implication is that some informative features have been thrown away or that some unwanted features have been retained.

3.4.5 A compression detection algorithm and worked example

Motivation: The previous discussion showed that compression induces changes to many of the quantities commonly used in data-driven process analyses. However, engineers are not always in a position to examine data closely enough to detect compression because plotting and examining the time trends is time consuming. If archived data are to be used for an automated analysis it is first necessary to test for the presence of compression.

If the number of spot values in the compressed archive and the original sampling rate are known then the compression factor may be determined by calculation as the ratio between the expected number of observations and the number of archived observations. However, such information is not always available when data sets are sent off-site to consultants or universities, and it may be necessary to estimate the compression factor from the reconstructed data only. An automated method for detection of piecewise linear compression is now presented and some guidelines given for its application to industrial data.

Compression detection procedure: Since the reconstructed data set is piecewise linear, its second derivative is zero everywhere apart from at the places where the linear segments join. Therefore the presence of the characteristic linear segments can be detected by counting of zero-valued second differences $\Delta(\Delta\hat{y})$ calculated from:

$$\Delta(\Delta\hat{y})_i = \frac{(\hat{y}_{i+1} - \hat{y}_i)/h - (\hat{y}_i - \hat{y}_{i-1})/h}{h} = \frac{\hat{y}_{i+1} - 2\hat{y}_i + \hat{y}_{i-1}}{h^2}$$

where $\hat{y}$ is the reconstructed signal and h is the sampling interval. The index i ranges from 2 to N-1, where N is the number of samples. Suppose the original data set had N values and after compression there are m archived spot values and $m-1$ linear segments. If the reconstructed data are differenced twice there will be $n = N - m$ second differences whose values are zero. Therefore the compression factor can be determined from:

$$CF_{est} = \frac{N}{m}$$

where $m = N - n$. For example, with ten data points compressed to four archived values and three linear segments there are $10 - 4 = 6$ second differences whose values are zero.

Implementation considerations: Enhancements are needed to the basic algorithm for industrial implementation for the following reasons:

- The sampling interval of the reconstructed signal may be larger than the original; e.g. the compression algorithm may use 10s samples but reconstruction may use 1 min samples.
- The effects of finite precision arithmetic mean that some computed second differences may not be exactly zero.

Suggestions for handling these cases are given here.

Dealing with a larger sampling interval: It is recommended to reconstruct compressed data using the same sampling interval as the original because reconstruction with a longer sampling interval leads to an underestimate of the compression factor. For instance, if five 10s samples out of 120 were archived then the true compression factor is 24 ($m=5, N=120$). When the data are reconstructed using 1 minute samples the number of piecewise linear segments does not change but there are only 20 samples in the reconstructed data so the compression factor appears to be 4.

An effect of reconstruction with a longer sampling interval is that the true end points of the piecewise linear segments may fall between samples. Thus x_i would be the end of one linear segment, x_{i+1} would be the start of the next with the true end point somewhere in between. The effect on the second differences is that there are two non-zero second differences where the linear segments join instead of the one that would be expected. The presence of these pairs of non-zero second differences can be used as a warning of a sampling interval issue. If such pairs are detected then the calculation of the compression factor has to acknowledge that each pair represents just one true archived point and the expression for the compression factor is modified to:

$$CF_{est} = \frac{N}{m/2}$$

Such pairs were detected in the industrial data of Figure 3.5 and therefore the modified expression was used in the calculation. The estimated compression factors are shown on the right of Figure 3.9. For instance, tag 20 has a compression factor of 41.7. It had 1428 zero second differences, 72 non-zero second derivatives in 36 pairs and 36 linear segments.

If the characteristic pairs are noticed then a warning must be given that the compression factors are under estimates and to reconstruct at the correct rate. Figure 3.9 shows such a warning.

Finite precision arithmetic: A procedure was developed to deal with the effects of finite precision and rounding errors. The numerical values of the second differences were converted to integers. The *ceil* function used in the following expressions rounds up to the next integer:

$$P = ceil\left(\log_{10}|x|\right)$$
$$y = x/10^{P}$$
$$z = y \times 10^{N}$$

x is the original entry in the data base having N significant figures, $P-1$ of which are to the left of the decimal point (e.g. $P=5$ and $N=10$ in 1478.144165). y has the same digits as x but has a zero to the left of the decimal point (e.g. 0.1478144165) and z is an integer with the same digits (e.g. 1478144165). The second difference calculations were applied to the integers z.

Some computed second differences may not be exactly zero because of arithmetic rounding errors. With the integer transformation above it would be expected that the errors would be ± 1. It was observed in the data set in Figure 3.4, however, that errors of up to ± 500 were present. That

is to say, the precision of the arithmetic used by the data historian in the reconstruction was less than 10 significant figures although the results were reported to 10 significant figures. The following sequence illustrates the pattern of second differences observed in z for a portion of a straight line trend in tag 7:

$$\left\{\begin{array}{l} -476,\ 477,\ 0,\ -477,\ 477,\ 0,\ -1,\ -476,\ 477, \\ \qquad 0,\ 0,\ -477,\ 477,\ 1,\ 477,\ -1,\ -476,\ 477,\ 0 \end{array}\right\}$$

Any second difference in z whose absolute value was below 500 was counted for calculation of the compression factor.

There is no fundamental significance to the numerical value of ± 477. The observed rounding errors arise from an interplay between the original data values, the arithmetic precision and the details of way the vendor has constructed the data base. It is system specific and therefore not possible to generalize. If the data historian complies with a published numerical Standard (e.g. IEEE 854-1987) then the threshold for second differences may be determined from the Standard. Otherwise the threshold must be determined by inspection of the arithmetical precision achieved and the number of significant figures in use, as was done here.

Compression detection in the industrial data set: Figure 3.9 shows compression factors estimated by the compression algorithm for the industrial data set of Figure 3.5. The algorithm detected the pairs of non-zero second differences that are characteristic of reconstruction at a different sampling interval than the original and has given a warning message. The compression factors are very large for many of the tags and the conclusion is that this data set is unsuitable for any data driven analysis.

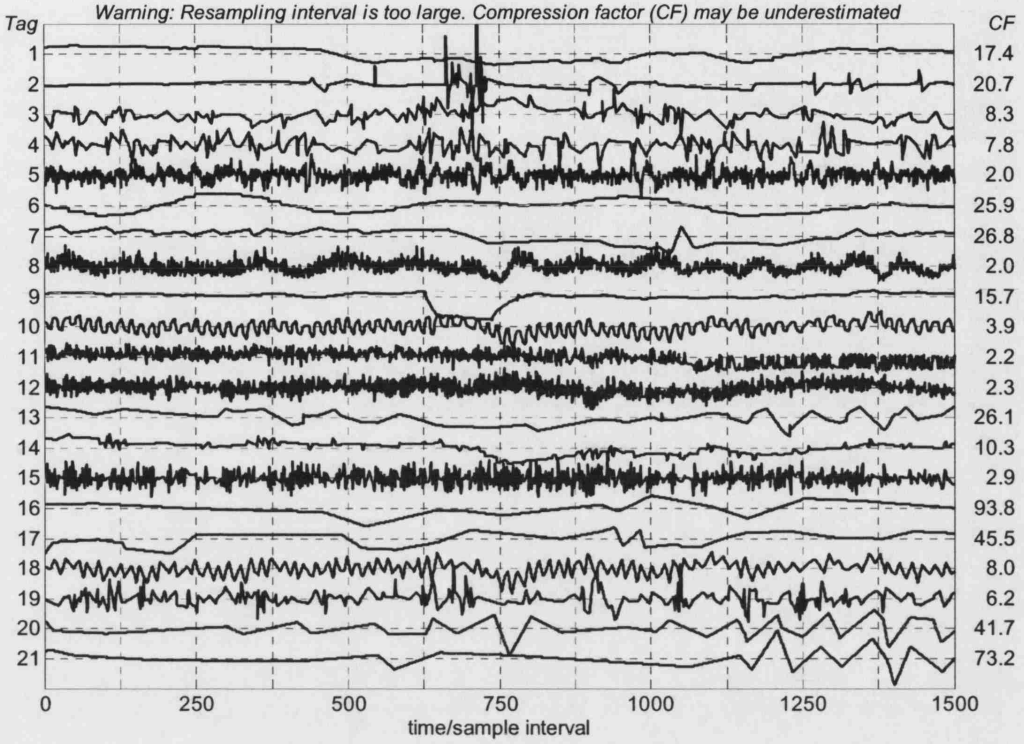

Fig 3.9 An industrial data set with compression in some tags showing an estimate of the CF and a warning that the data are being reconstructed at a different sampling rate than the original.

3.5 CHAPTER SUMMARY

The chapter has described and discussed several procedures that should be applied before a data-driven analysis begins.

High density plots provide a means of visualizing the time trends and useful transforms in the form of power spectra and autocovariance functions. The high density plots mean centre and scale the data to unit standard deviation and hence remove important statistical information. The original means and variances are captured in a table of basic statistics which can be considered alongside the high density plots.

Frequency domain filtering gives a process control engineer the opportunity to refine the data and to zoom-in for a close look at selected frequency ranges. This chapter has presented the detailed formulation of the filter and given an example of its use. It has been applied at several other places in the thesis and an automated way of selecting the filter boundaries for the purposes of detection of multiple oscillations will be given in Chapter 4.

A strong recommendation from this thesis is that process operating companies should not use data compression when archiving process data. This chapter has shown that data compression using the popular swinging door algorithm destroys key information. Even the basic statistics (mean and standard deviation) are changed by compression and compression renders the data worthless for data-driven analysis. An algorithm for the automated detection of compression was presented and demonstrated.

4. OSCILLATION DETECTION

4.1 OVERVIEW

4.1.1 Motivation

For process control engineers engaged in plant maintenance the detection of oscillating control loops is a basic activity because oscillations cause loss of profits. Features of interest are:

- Detection of the presence of one or more oscillations indicated by a regular pattern in the data;
- Determination of the periods of the oscillations, which can be used to infer the presence of a plant-wide oscillation;
- Determination of the magnitudes of the oscillations. A large magnitude oscillation needs to be addressed because it increases process variability and reduces profitability.

The aim of the oscillation detection tools is to detect distributed oscillations, to find the periods of the oscillations and determine which tags are affected by which oscillation.

4.1.2 Introduction

Chapter 4 describes software tools for the purposes of detection of distributed oscillations. They are illustrated with a case study from a SE Asian refinery which is introduced in section 4.2.

The oscillation detection algorithms are then described in detail. The focus is on the detection of groups of similar oscillations in multiple plant measurements that indicate the presence of a distributed oscillation, and the thesis thus extends earlier single-loop work in the literature. The case where more than one oscillation is present in a single time trend has also been solved. The estimation of the period of oscillation is based on the intervals between zero crossings of the autocovariance function. A proof is given in Appendix A2 which shows the zero crossings method is a fundamentally more accurate way to determine the period of an oscillation that the alternative of taking the reciprocal of the peak frequency in the power spectrum.

Section 4.3 gives results of an automated procedure for oscillation detection in which three oscillatory disturbances were found.

4.1.3 Work done by the thesis author and others

Work reported in this chapter has been published in the following paper:

> Thornhill, N.F., Huang, B., and Zhang, H., 2003, Detection of multiple oscillations in control loops, *Journal of Process Control,* **13**, 91-100.

The paper was written by Thornhill who also created all software, results, diagrams and graphs while the co-authors read, discussed and criticized the text and helped to give an industrial relevance.

4.2 SE ASIA REFINERY CASE STUDY

4.2.1 Process description

Ownership of the case study: The techniques described in this chapter will be illustrated using a data set that was originally examined by Dr Anand Vishnubhotla of Matrikon Inc, Edmonton, Canada. The owner of the data was a refinery in SE Asia. Permission to publish the data and schematic was requested and gained from the owners by Thornhill. Figure 4.1 shows a simplified process schematic provided by the refinery. The measurements are from 37 process tags of a hydrogen reformer.

Process description: The operation of a reformer is as follows (Terrible *et.al.*, 1999). The aim of the process is to make hydrogen from hydrocarbons. The mixed feed includes propane, liquid petroleum gas, naphtha and tail gases (waste gas) of varying composition from other units in the refinery. Pre-treatment in the processing block labelled *feed vaporiser/preheat* compresses the mixed gases. It is then mixed with steam and heated to reaction temperature in excess of 800C. Two main reactions occur in the reformer and the end result is an equilibrium mixture of hydrogen, methane, carbon dioxide and carbon monoxide. The reactions below show a hydrocarbon reacting with steam to give carbon monoxide and hydrogen, and a reversible reaction called the *water gas shift reaction* in which carbon monoxide reacts with steam to give hydrogen and carbon dioxide. The reaction uses a nickel catalyst and conditions are set up to favour the forward reaction to maximise the production of hydrogen:

$$C_nH_{2m} + nH_2O \rightarrow (m+n)H_2 + nCO$$

$$CO + H_2O \leftrightarrow H_2 + CO_2$$

The last stage of the process is a pressure swing adsorption unit (PSA) in which methane, carbon dioxide and carbon monoxide are removed to give a high purity hydrogen. Its off-gas contains methane which can be utilized as a fuel and which is recycled to the reformer.

4.2.2 Process data

Description of data: Data were sampled at 1 minute intervals from the plant for about 8.5 hours of running and Figure 4.2 shows a 500 minute window of the time trends of the process measurements and controller errors. Mean centred controller errors are the negative of the process measurement apart from in the case of controllers such as FC1, FC4, FC5 and FC8 which are in a cascade configuration with a changing set point. Indicators such as FI1 (Tag 34) are treated for the purposes of determining controller error as a controller with set point equal to the mean value.

Distributed oscillations: Some clusters of oscillating measurements can be observed by eye, for instance the trends for tags 2-4, 10, 11, 19, 20, 33 and 34 are all characterized by the same oscillation, while tags 35 and 36 also look similar to one another. The challenge for the oscillation detection techniques of this chapter is to detect those groups of tags having similar oscillations and find any other members of the clusters that are not so easily detected by eye.

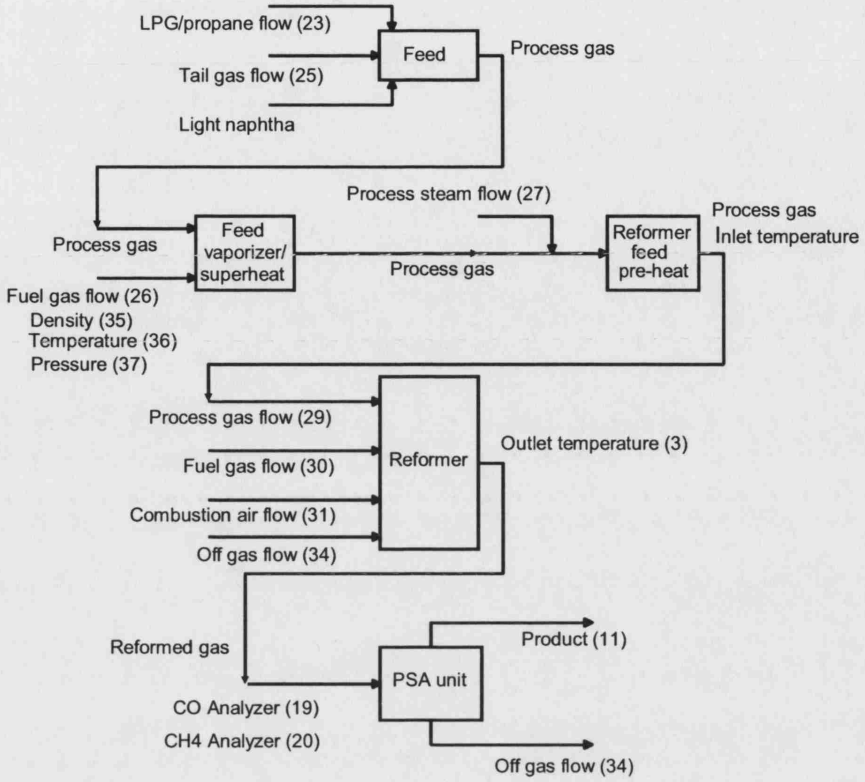

Fig 4.1 Process schematic of SE Asia hydrogen plant.

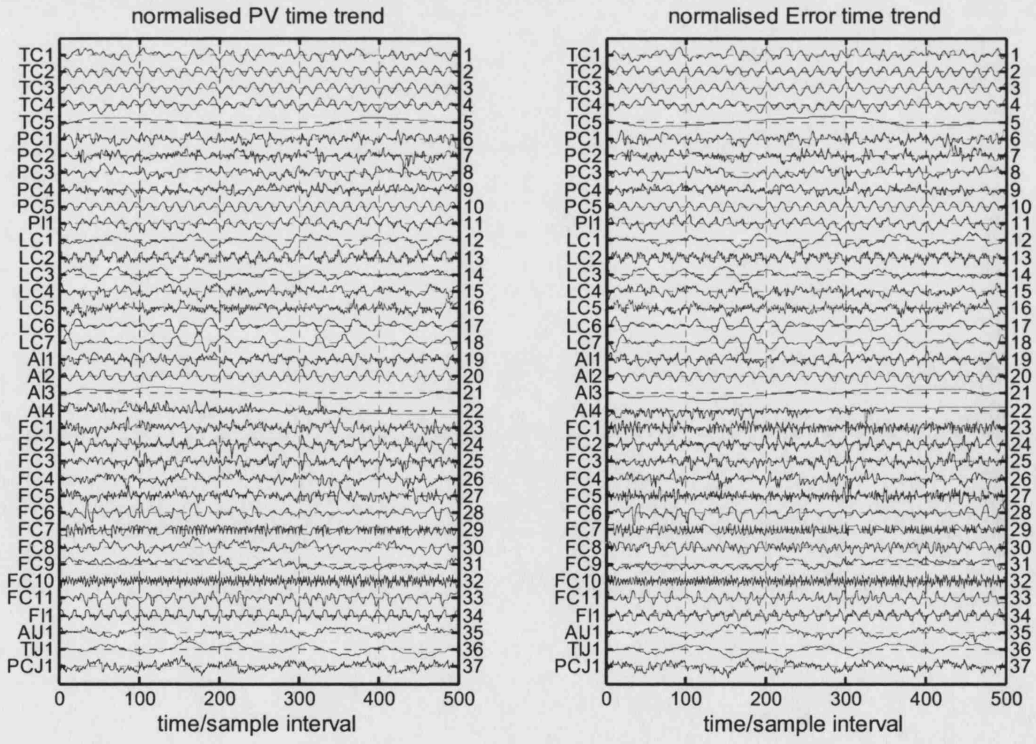

Fig 4.2 Time trends of SE Asia data. Left: Process variables Right: Controller errors.

4.3 DETECTION OF OSCILLATIONS

4.3.1 Introduction to oscillation detection

Overview and background: The basis of the oscillation detection method is the finding of the zero-crossings of an oscillating signal. If the zero crossings are regularly spaced then an oscillation is present. The key steps are detection of the zero-crossings and a decision about their regularity.

Several authors have addressed the detection of oscillatory measurements in process data. Based on time trend data, Hägglund (1995) detected zero crossings of the controller error signal in a control loop and calculated the integrated absolute error (IAE) between successive zero crossings. An oscillatory time trend has larger IAE values than a random one. This was an important step forward and allowed a real-time implementation whereby oscillations in a control loop could be detected on-line. Thornhill and Hägglund (1997) inspected the regularity of the zero crossings of a time trend with a view to an off-line analysis of individual control loops, for example an audit of historical data from the previous day's running of the plant. The presence of an oscillation was inferred when the regularity was high and an estimate was given of the signal to noise ratio of the oscillation.

Contributions of the thesis: The work reported in this chapter extends the work of the previous literature in two ways. Firstly, it uses zero-crossings of the autocovariance function for oscillation detection rather than zero crossings of the time series. As discussed in Chapter 2, the autocovariance function of an oscillating signal is itself oscillatory with the same period as the oscillation in the time trend. Oscillations can therefore be detected if the pattern of zero crossings of the autocovariance function is regular. The benefit of the use of autocovariances is that they are much less noisy than the time trend data and the problem of spurious zero crossings reported in Thornhill and Hägglund (1997) therefore goes away. For instance, in the case of white noise all the noise appears in the zero lag channel of the autocovariance. Therefore noise does not disrupt the zero crossing of the autocovariance function in the same way as it disrupts zero crossings in the time domain, and the pattern of zero crossings reveals the presence of an oscillation more clearly than the zero crossings of the time trend. Figure 4.3 demonstrates that the pattern of zero crossings of the *ACF* reveals the presence of an oscillation more clearly than the zero crossings of the time trend.

The second major extension has been in the detection of plant-wide oscillations. In contrast to previous work where oscillation was detected only within an individual control loop, the work reported here finds all oscillations present in all the tags and reports groups of tags that are being influenced by the same oscillation. This extension significantly aids diagnosis. It is impossible to tell from analysis of a single control loop whether the oscillation originates within the loop or is

caused by an external disturbance. However, if a groups of tags has been identified as sharing the same oscillation then it is reasonable to suppose that there is just one root cause to be found and that the remaining oscillations are secondary propagated disturbance. The analysis therefore greatly aids the maintenance task by revealing the true number of oscillatory faults present in the plant.

A proof is presented in Appendix A2 that shows estimates of oscillation period from zero crossings are more accurate than those determined from taking the reciprocal of the frequency of the main peak in the power spectrum, thus validating the choice of the zero-crossings method.

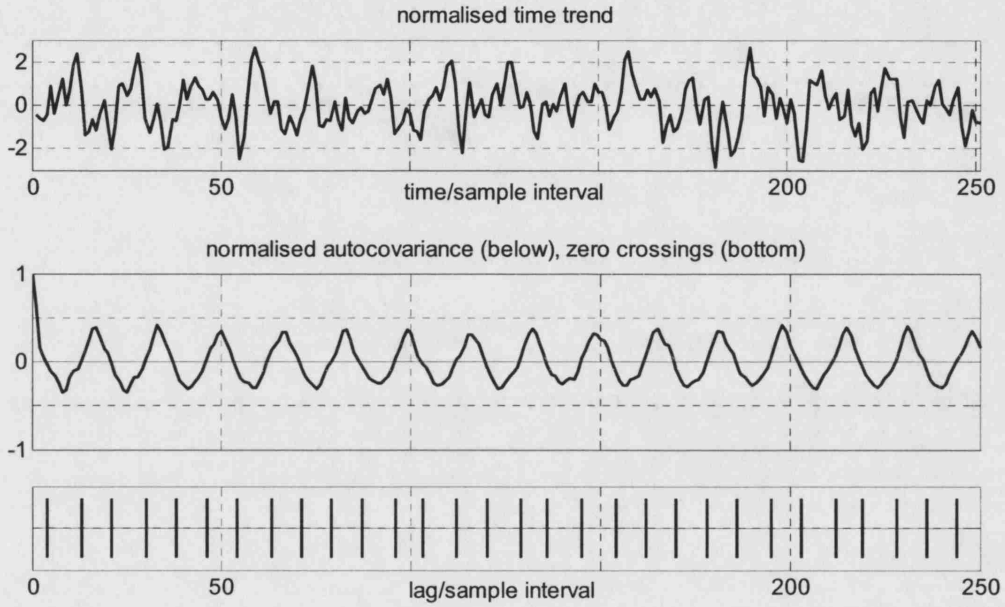

Fig 4.3 Oscillation detection in tag 25 of the SE Asia data set. Upper plot: Portion of time trend Middle: Autocovariance function. Bottom: Zero crossings.

4.3.2 Techniques

Assessment of period and regularity of oscillation: Each oscillation has two zero crossings and hence the intervals between zero crossings are:

$$interval = \frac{1}{2}\left(\overline{T_p} \pm \Delta T_p\right)$$

where $\overline{T_p}$ is the mean period and ΔT_p a random variation in the period. Thus $\overline{T_p}$ is twice the mean value of the intervals and the standard deviation of the period is $\sigma_{T_p} = 2 \times \sigma_{intervals}$ where $\sigma_{intervals}$ is the standard deviation of the intervals.

The ten intervals between the first eleven zero crossings are used for calculation of $\overline{T_p}$ and σ_{T_p}. The interval from lag zero up to the first zero crossing is excluded from the calculation because it corresponds to only one half of a completed deviation. Ten intervals are used because except in the case of a very persistent oscillation the magnitudes of the autocovariance functions tend to have decaying profiles and some spurious zero crossings may be found at large lags. Fewer intervals (e.g. eight) can be used for the regularity assessment when the period of oscillation is long but it is not recommended to use fewer than four intervals because the estimates of $\overline{T_p}$ and σ_{T_p} become unreliable.

An oscillation is considered to be regular if the standard deviation of the period is less than one third of the mean value. A regularity statistic is defined where values of $r > 1$ are taken to indicate a regular oscillation with a well defined period:

$$r = \frac{1}{3} \times \frac{\overline{T_p}}{\sigma_{T_p}}$$

The threshold value for r has a statistical interpretation. If the zero crossings were random with an equal probability of arrival in each sampling interval then the intervals between zero crossings would have an exponential distribution (Freund, 1971):

$$f(T_p) = \alpha e^{-\alpha T_p}$$

The standard deviation of an exponential distribution is equal to its mean and both are equal to $1/\alpha$ so the null hypothesis of random arrivals would be $\overline{T_p} = \sigma_{T_p}$. The observed σ_{T_p} value gives an estimate of $1/\alpha$ and the observed value of $\overline{T_p}$ is then tested to see if it could plausibly have been generated by the exponential distribution $\alpha e^{-\alpha T_p}$. The null hypothesis is rejected if:

$$\overline{T_p} > 3 \times 1/\alpha \qquad \text{or} \qquad \overline{T_p} > 3\sigma_{T_p}$$

The statistic $\overline{T_p} = \frac{1}{N}\sum_{i=1}^{N} T_{p_i}$ is constructed from observations T_{p_i} of the sampling interval. It is a statistical quantity because each different set of N random observations would give a slightly different value of $\overline{T_p}$. It can be shown that $E(\overline{T_p}) = 1/\alpha$ where E is the expectation operator, and the estimate is therefore unbiased. The variance of $\overline{T_p}$ under the null hypothesis is:

$$Var(\overline{T_p}) = E\left(\overline{T_p} - E(\overline{T_p})\right)^2 = E\left(\overline{T_p} - \frac{1}{\alpha}\right)^2 = \frac{1}{N\alpha^2}$$

and its standard deviation under the null hypothesis is $Std(\overline{T_p}) = 1/\sqrt{N}\alpha$.

The regularity index test for rejection of the null hypothesis $\overline{T_p} > 3 \times 1/\alpha$ implies that the distance of the observed $\overline{T_p}$ from the expected value is:

$$\overline{T_p} - E(\overline{T_p}) = 2 \times \frac{1}{\alpha} = 2\sqrt{N} Std(\overline{T_p})$$

Measurements with more than one oscillation: The zero crossings of the *ACF* may not be regular if more than one oscillation is present. Figure 4.4 shows a portion of time trend and *ACF* from an industrial plant measurement having two superimposed oscillation of different periods (it is from the case study of Chapter 8). The bottom panel in the figure marks the positions of the zero crossings. The intervals between zero crossings of the autocovariance function reflect neither oscillation accurately because the zero crossings of the fast and slow oscillations each destroy the regularity of the other's pattern.

The problem is addressed by frequency domain filtering (Section 3.3). A filtered *ACF* is calculated from the inverse Fourier transform of the filtered power spectrum. The frequency-domain filter sets the power in unwanted frequency channels to zero. Figure 4.5 shows the example data from Figure 4.4 after filtering to separate the two oscillations at about 330 samples per cycle and 65 samples per cycle. The filter ranges were 100 to 1000 sample per cycle and 35 to 100 samples per cycle respectively. By removal of the interfering oscillation, the filtering makes the zero crossings more regular and therefore suitable for use in oscillation detection.

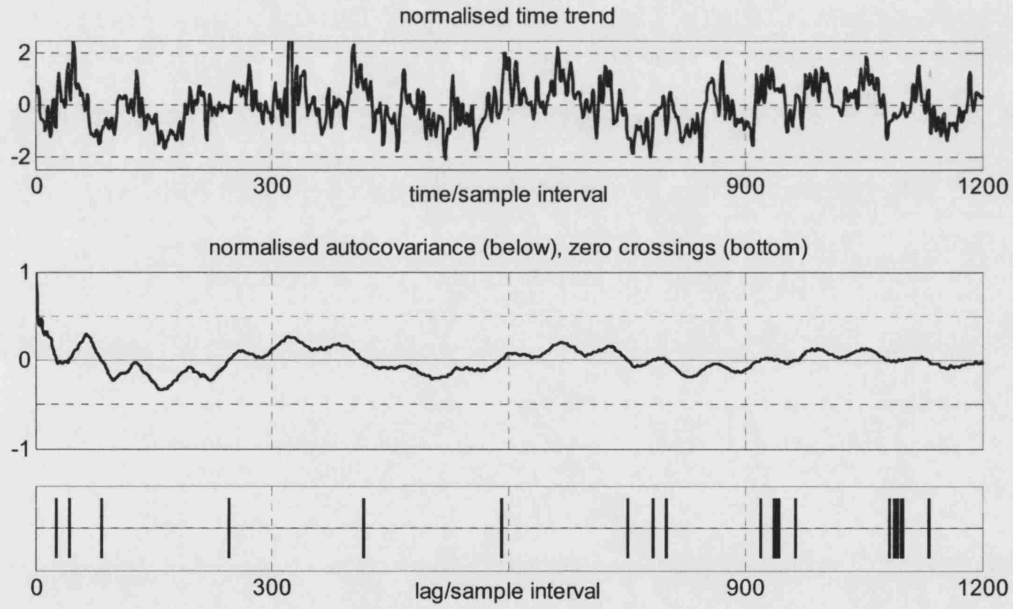

Fig 4.4 Oscillation detection in a tag with multiple oscillations. The zero crossings are not regular. Upper plot: Time trend Middle: Autocovariance Bottom: Zero crossings.

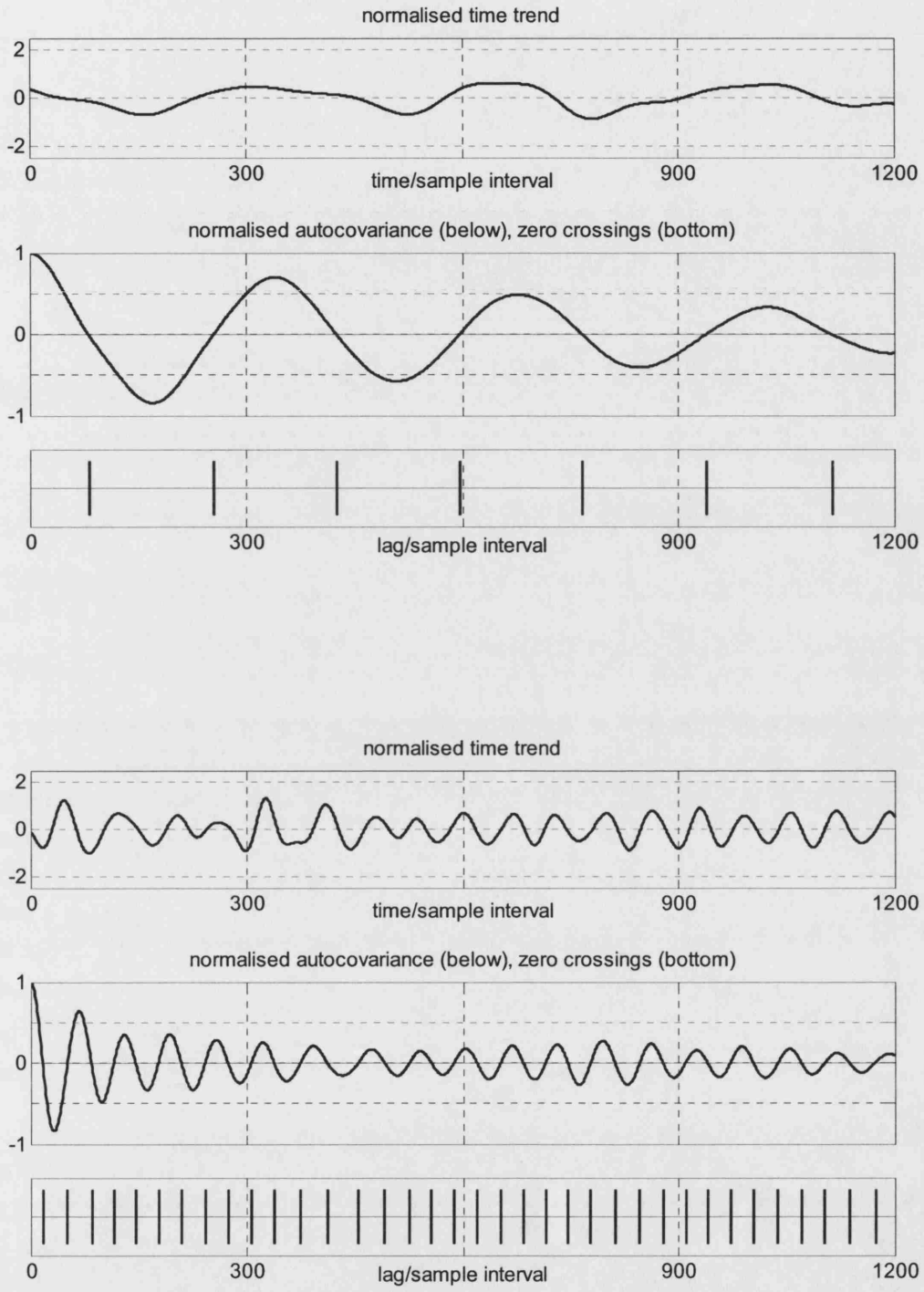

Fig 4.5 Oscillation detection using filtered data from Fig 4.4. Upper three panels: data filtered to retain the long period oscillation at about 330 samples per cycle. Lower three panels: data filtered to retain the fast oscillation at about 65 samples per cycle.

Assessment of the strength of an oscillation: The power spectrum is normalized before filtering, therefore $\sum \hat{P}_i = 1$, where $\hat{P}_i$ is the normalized power in the i'th frequency channel. Thus, when the spectrum is filtered it follows that the summed spectral power in the selected frequency channels gives the power as a fraction of the total. These power values include both oscillatory components and noise. However, if there is a regular oscillation having $r > 1$ the summed power in the filtered spectrum would be dominated by the power of the oscillatory signal and may be utilized as an approximation to the power of the oscillation.

4.3.3 Automation

Characterization of the plant-wide oscillations: The oscillation detection tool returns an array of oscillation statistics. They are:

- Mean period of oscillation, $\overline{T_p}$;
- Standard deviation of the period, σ_{T_p};
- Oscillation index, r;
- Power in the oscillation expressed as a percentage.

Assignment of oscillations to groups: Each detected oscillation is assigned to a group of similar oscillations and the group is characterized by its most regular member. The condition for membership of the group is that the period of oscillation of the tag under test should be no more different than one standard deviation from the period of the most regular oscillation in the group. The routine also screens out tags with low signal strength. Those having less than 10% spectral power in the frequency range of the oscillation are excluded from the group.

The method used for automated counting and grouping of detected oscillations was adapted from a classification procedure by Chatfield and Collins (1980).

Algorithm: Automated grouping of oscillations

Step 1. The periods T_p of the significant oscillations (those with $r \geq 1.0$ and power above 10%) are arranged in a vector. The standard deviations of the periods σ are arranged in another vector and their oscillation indexes, r, in a third. A fourth vector, w, of the same length as the others is initialized to all ones.

Step 2: Scaled differences $d_{i,j}$ between the periods of oscillation are calculated:

$$d_{i,j} = \frac{\left| T_{p_i} - T_{p_j} \right|}{\max(\sigma_i, \sigma_j)}$$

The condition $d_{i,j} < 1$ indicates that two oscillations belong together because the difference between the periods of two oscillations is less than the larger of the two standard deviations.

The $d_{i,j}$ are arranged into a symmetric matrix. The row and column indexes (i and j) for the smallest of any $d_{i,j} < 1$ are determined. Only the upper off-diagonal elements are considered because the diagonal elements are zero and the lower off-diagonal elements are the same as the upper elements.

Step 3: The vectors are revised. T_{p_j} is deleted from the T_p – vector and the more regular of the two oscillations is recorded. i.e. T_{p_i} is kept if $r_i > r_j$ or replaced by T_{p_j} if $r_j > r_i$. Likewise, σ_j is deleted from the σ – vector and σ_i is kept if $r_i > r_j$ or replaced by σ_j if $r_j > r_i$. In the r – vector, r_i is replaced by $\max(r_i, r_j)$ and r_j is deleted. In the w – vector, w_i is replaced by $w_i + w_j$ and w_j is deleted. Thus the w – vector keeps track of how many oscillations have been grouped.

Step 4: Steps 2 and 3 are repeated until there are no more changes in the vector of T_p values (i.e. the $d_{i,j}$ are all greater than one). The final result is a short T_p vector giving the period of the most regular oscillation in each group of oscillations and a corresponding w vector giving the number of oscillations in the group.

Automated detection of oscillations in different frequency ranges: An automated means has been devised of choosing filter ranges for whole data sets with multiple oscillations at different frequencies. The basis of the method is to seek carefully for oscillations in all the tags whenever any oscillation was detected in one tag. Thus, for instance, if tag 30 showed an oscillation without any filtering at, say, 20 samples per cycle, then a focused search with filtering would commence over the whole data set using a filter centred on 20 samples per cycle.

The filter ranges may be designed using the following algorithmic approach.

Algorithm: Automated selection of frequency filter boundaries

Step 1: Apply oscillation detection to the pre-processed data. Record the periods of any oscillations detected using the automated grouping algorithm;

Step 2: Place filter boundaries mid-way on the frequency axis between any detected oscillations, and mid way between the highest (lowest) frequency oscillation and the highest (lowest) frequency in the pre-processing filter;

Step 3: Re-do the oscillation analysis for each new frequency range. Repeat step 3 for any frequency ranges where an oscillation at more than one frequency is present to uncover cases of multiple oscillations in the same tag. Stop sub-dividing when the range contains no multiple oscillations, or when the filters become too narrow (Section 3.3.3)

Step 4: Oscillation detection can generate spurious results when the filter boundaries are placed on the frequency axis such that they split a broad spectral feature across two ranges. The remedy for false detection is to ignore detected oscillations having periods that are within 10% of the filter boundary. For instance, if the filter were [35 200] then any oscillations at 35-38 samples per cycle or 180-200 samples per cycle should be disregarded.

4.4 WORKED EXAMPLE IN OSCILLATION DETECTION

4.4.1 Basic oscillation detection

Results without filtering: The first step in an industrial case study is to apply the oscillation detection technique to the data set without any frequency domain filtering. Oscillation detection without filtering is a blunt instrument because it will not detect tags having more than one oscillation present, nor will it detect oscillations in tags with non-stationery behaviour such as those subject to diurnal temperature variations. This section is illustrated with oscillation detection for the SE Asia data set. It shows that unfiltered data can nevertheless produce interesting and useful results and thus motivate a more detailed analysis. This section then shows the improvements to be gained from a more advanced analysis in which filtering is used to analyze tags with multiple oscillations.

Autocovariances and spectra: Figure 4.6 shows high density plots of spectra, autocovariance functions and their zero crossings and the raw results of oscillation analysis in the form of oscillation indexes. The indexes are the numbers labelled as r on the right hand side of the middle panel. An oscillation index above 1 ($r > 1$) indicates a regular oscillation. The percentage power is 100% for every tag because no filtering was used, therefore all frequencies are included.

In cases where single oscillations are present, for instance in TC2, TC3 and TC4 (Tags 2-4) the oscillations have been detected accurately and the oscillation indexes are high. However, without filtering it is not possible to be sure that no oscillation is present in other tags. For instance, a multiple oscillation is certainly present in PCJ1 (Tag 37) as can be readily seen by visual inspection of the autocovariance plot.

Slowly varying tags (5 and 21) had only one zero crossing and thus no period of oscillation and no oscillation index could be determined for these.

Oscillation statistics and report: The plant-wide oscillation reports for the SE Asia data are shown in Tables 4.1 and 4.2. The number of significant figures reported in the table follows the convention described in Chapter 3 in which the standard deviation is reported to two significant figures and the number of significant figures in the mean value adjusted to match. Hence, for instance, the raw result for tag 4 of 16.22 ± 2.11 is reported as 16.2 ± 2.1.

Table 4.2 shows there is a distributed plant-wide disturbance with a period of 16.4 samples per cycle involving 15 tags. This is the oscillation with a spectral peak at 0.06 on the normalized frequency axis that can be seen in the spectra in the right hand panel of Figure 4.6. A second oscillation involving three tags was found at 85.7 samples per cycle. The oscillation analysis has shown, additionally, that there is a single measurement (tag 14) with an oscillation period of about 50 samples per cycle and that tag 32 was also alone with a unique period of 3.6 samples per cycle.

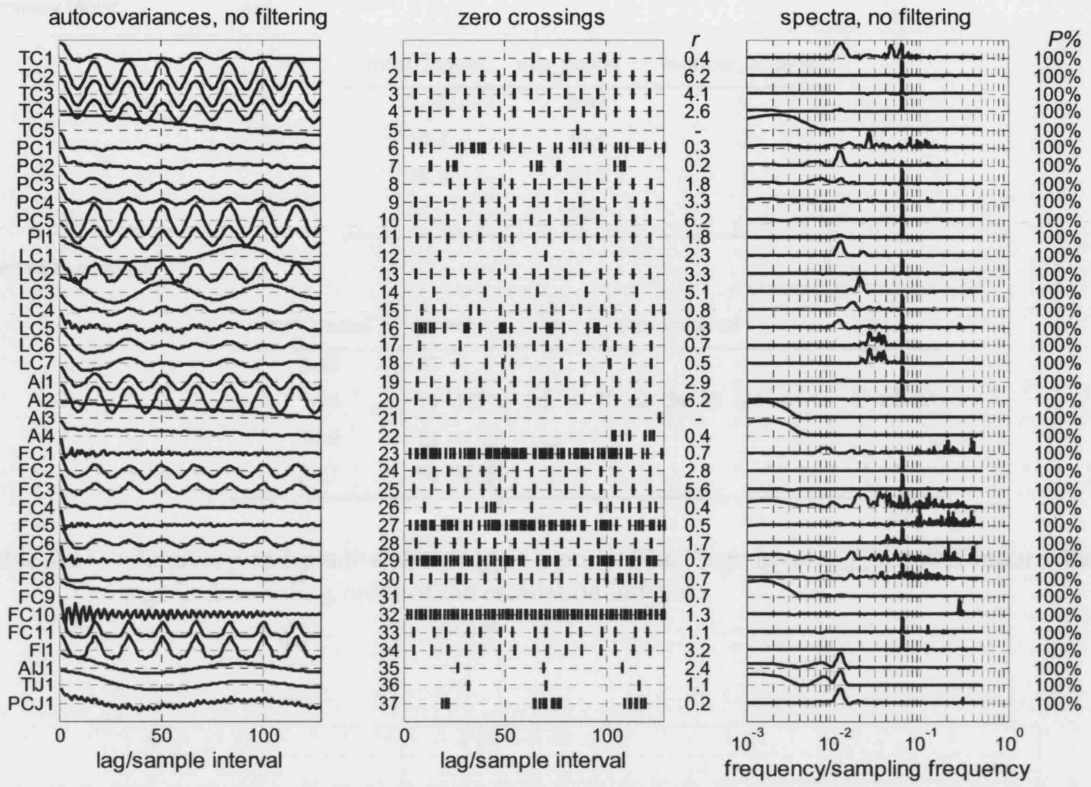

Fig 4.6 Plant-wide oscillation detection for SE Asia data. Any tag with an oscillation index (r)above 1 has regular zero crossings and is oscillatory.

tag no	$\overline{T_p}$	σ_{T_p}	r	power	tag no	$\overline{T_p}$	σ_{T_p}	r	power
1	28	22	0.41	100%	20	16.44	0.88	6.2	100%
2	16.44	0.88	6.2	100%	21	-	-	-	100%
3	16.4	1.3	4.1	100%	22	6.7	5	0.44	100%
4	16.2	2.1	2.6	100%	23	2.9	1.5	0.66	100%
5	-	-	-	100%	24	16.7	2	2.8	100%
6	7.8	7.5	0.35	100%	25	16.67	1	5.6	100%
7	14	23	0.21	100%	26	20	16	0.43	100%
8	16.2	3.1	1.8	100%	27	3.6	2.2	0.54	100%
9	16.4	1.7	3.3	100%	28	16	3.2	1.7	100%
10	16.44	0.88	6.2	100%	29	2.9	1.5	0.66	100%
11	16.2	3.1	1.8	100%	30	13.1	6.3	0.69	100%
12	86	12	2.3	100%	31	12	5.7	0.71	100%
13	16.4	1.7	3.3	100%	32	3.56	0.88	1.3	100%
14	50.9	3.3	5.1	100%	33	16.9	5.2	1.1	100%
15	15.6	6.3	0.82	100%	34	16.7	1.7	3.2	100%
16	5.8	5.9	0.33	100%	35	86	12	2.4	100%
17	20.2	9.1	0.74	100%	36	85	26	1.1	100%
18	23	14	0.55	100%	37	12	27	0.15	100%
19	16.2	1.9	2.9	100%					

Table 4.1 Plant-wide oscillation analysis for the SE Asia data.

Summary table:

most_regular	ave_period	number_of_tags
3.56	3.56	1
16.4	16.4	15
50.9	50.9	1
85.6	85.7	3

Plant-wide analysis

ave_period	tags involved
3.56	32
16.4	10 20 2 25 3 13 9 34 19 24 4 11 8 28 33
50.9	14
85.7	35 12 36

Table 4.2 Summary and plant-wide analysis reports. The "tags involved" column reports the tags in descending order of the oscillation indexes.

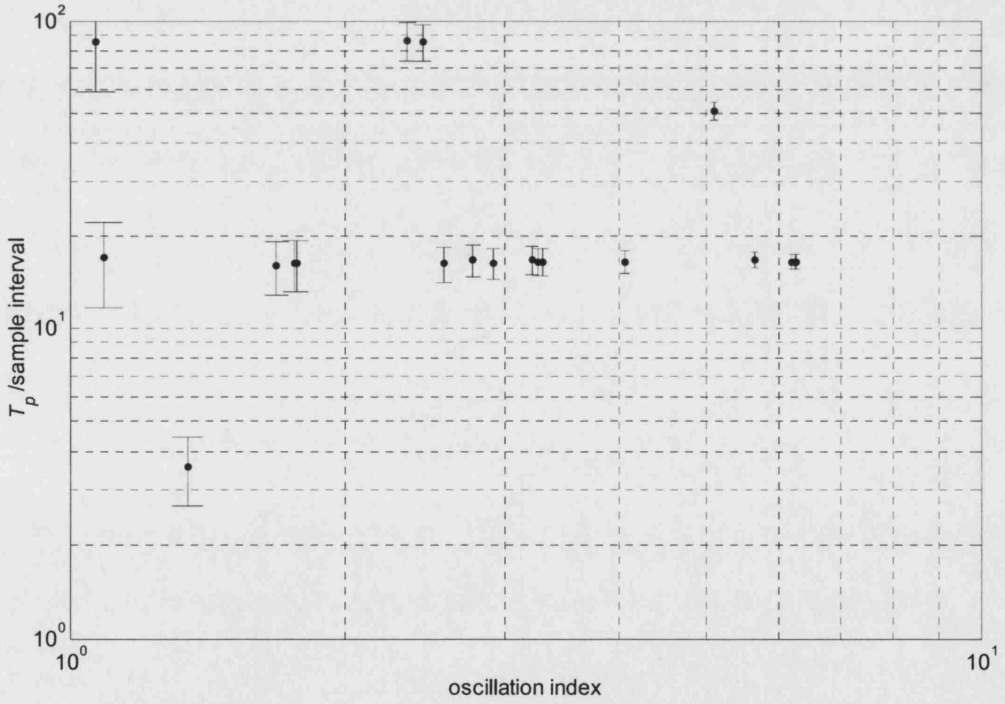

Fig 4.7 Graphical representation of plant-wide analysis. Each spot represents one tag showing its oscillation period (vertical axis) and oscillation index (horizontal axis). The error bars show σ_{T_p}. A large group of tags is oscillating with a period of 16.4 samples per cycle.

4.4.2 Multiple oscillation detection

Filter design: Since the SE Asia data set had four distinct oscillations present it is necessary to apply filtering to determine whether any of the tags are affected by more than one oscillation.

The periods of oscillation were: 3.56, 16.4, 50.9 and 85.7 samples per cycle, equivalent to 0.281, 0.0610, 0.0196 and 0.0117 cycles per sample on the normalized frequency axis. Also, the lowest and highest frequencies present in the unfiltered data set were 2 samples per cycle (the Nyquist frequency) and 512 samples per cycle (the length of the data set), equivalent to normalized frequencies of 0.5 and 0.0019.

Following the algorithm for automated selection of frequency filter boundaries, the filter boundaries are placed at the mid-point between the oscillation frequencies at 0.171, 0.0403 and 0.0157 on the normalized frequency axis. To the nearest integer, the samples per cycle boundaries for the filters are given by the reciprocals of these frequencies as follows: [6 to 25], [25 to 64]. These filters embrace the oscillations at 16.4 and 50.9 samples per cycle.

The algorithm also calls for additional filters for the fastest and slowest oscillations at 3.56 and 85.7 samples per cycle. The data set had 512 samples so the lowest normalised frequency present was 0.0019. The mid-point filter boundary is therefore 0.0068 or 147 samples per cycle and the range for the filter is [64 to 147] samples per cycle. The highest frequency is 0.5 or two samples per cycle. The filter for the fast oscillation is [3 to 6] samples per cycle.

The table below sets out the above calculations in an ordered form:

	samples/cycle	normalised frequency	mid-point frequency	filter boundary samples/cycle
highest frequency	2	0.5		
			0.390	3
oscillation	3.56	0.281		
			0.171	6
oscillation	16.4	0.0610		
			0.0403	25
oscillation	50.9	0.0196		
			0.0157	64
oscillation	85.7	0.0117		
			0.0068	147
lowest frequency	512	0.0019		

Table 4.3 Calculation table for the frequency-domain filter design.

Results with filtering: The tables of results with the various filters are presented below.

Filtering [3 to 6] samples per cycle: The filter has detected additional tags affected by the fast oscillation with 3.56 sample per cycle. The table also shows results in italics for a new oscillation at 5.4 samples per cycle but this is disregarded because it is very close to the filter boundary of 6 samples per cycle, as explained in section 4.3.3.

Summary table:

most_regular	ave_period	number_of_tags
3.56	3.56	7
6	*5.4*	*7*

Plant-wide analysis

ave_period	tags involved
3.55	32 16 13 19 37 15 34

Table 4.4 Summary and plant-wide analysis reports for oscillation detection in the range [3 to 6] samples per cycle. The oscillation with a period of 5.4 is disregarded because it is close to the filter boundary.

Filtering [6 to 25] samples per cycle:

Summary table:

most_regular	ave_period	number_of_tags
12.7	12.7	1
16.4	16.8	23

Plant-wide analysis

ave_period	tags involved
12.7	22
16.8	10 4 3 20 11 2 16 24 25 17 15 8 31 13 19 18 9 34 12 28 37 1 33

Table 4.5 Summary and plant-wide analysis reports for oscillation detection in the range [6 to 25] samples per cycle.

The [6 to 25] filter detected additional tags affected by the fast oscillation with 16.4 sample per cycle and has adjusted its estimate of the average period to 16.8 samples per cycle. In total 23 out of 37 tags are affected by this oscillation and it truly is a serious plant-wide disturbance. Its origin is in the off-gas recycle from the PSA unit to the reformer and will be discussed in more detail in Chapter 6. The table also shows a previously undetected oscillation at 12.7 samples per cycle in tag 22. A spurious oscillation close to the filter boundary of 6 samples per cycle was disregarded and is not reported in the table. Plots of the filtered autocovariance functions and spectra are shown in Figure 4.8 and discussed in the next sub-section.

Filtering [25 to 64] samples per cycle:

Summary table:

most_regular	ave_period	number_of_tags
33.3	32.3	2
39.3	38.0	3
46.0	46.0	1
50.9	54.3	3

Plant-wide analysis

ave_period	tags involved
32.3	9 17
38.0	6 16 1
46.0	12
54.3	14 37 15

Table 4.6 Summary and plant-wide analysis reports for oscillation detection in the range [25 to 64] samples per cycle.

The [25 to 64] filter detected two additional tags besides tag 14 affected by the oscillation with 50.9 samples per cycle and has adjusted its estimate of the average period to 54.3 samples per cycle. An oscillation was detected in tags 12 at 46 samples per cycle. Two other new oscillations were detected at 32.3 and 38.0 samples per cycle. These are local rather than plant-wide because only a few tags are involved.

Filtering [64 to 147] samples per cycle:

Summary table:

most_regular	ave_period	number_of_tags
85.2	84.84	11

Plant-wide analysis

ave_period	tags involved
84.8	12 37 1 16 4 11 19 7 35 36 21

Table 4.7 Summary and plant-wide analysis reports for oscillation detection in the range [64 to 147] samples per cycle.

Analysis of the unfiltered data previously found three tags (35, 12 and 36) oscillating with an average period of 85.7 samples per cycle. The [64 to 147] filter removes other interfering oscillation and shows a total of 11 tags oscillating in that group with an average period of 84.8 samples per cycle. This is therefore a plant-wide disturbance. Its origin is in the thermodynamic properties of fuel gas that provides pre-heat for the reformer.

Autocovariances and spectra: Figure 4.8 examines the impact of filtering on the autocovariance functions and spectra illustrating the results from the [6 to 25] samples per cycle filter. Compared to Figure 4.6, it can be seen that the effect on the spectra (right hand panel) has been to remove both low and very high frequencies and to focus upon the range which contains the 16.4 sample per cycle oscillation (about 0.06 on the frequency axis). More of the autocovariance functions are now oscillating at or close to 16.4 samples per cycle. For instance, Figure 4.8 shows that tags 15 to 18 (LC4 to LC7) have oscillatory autocovariance, that their oscillation indexes are high and that the power in the oscillation is high. Therefore filtering has revealed that these tags are affected by the plant-wide disturbance whereas in Figure 4.6 the patterns of their zero crossings were irregular and they were not identified.

It can be seen in tables 4.4 to 4.7 the level loops 15, 16, 17 and 18 are affected by many of the disturbances in the plant. It is likely, therefore, that these particular level controllers are configured as surge tanks and are slackly tuned expressly for purpose of buffering of disturbances. The purpose of a surge tank is to catch and smooth out disturbances in liquid flows.

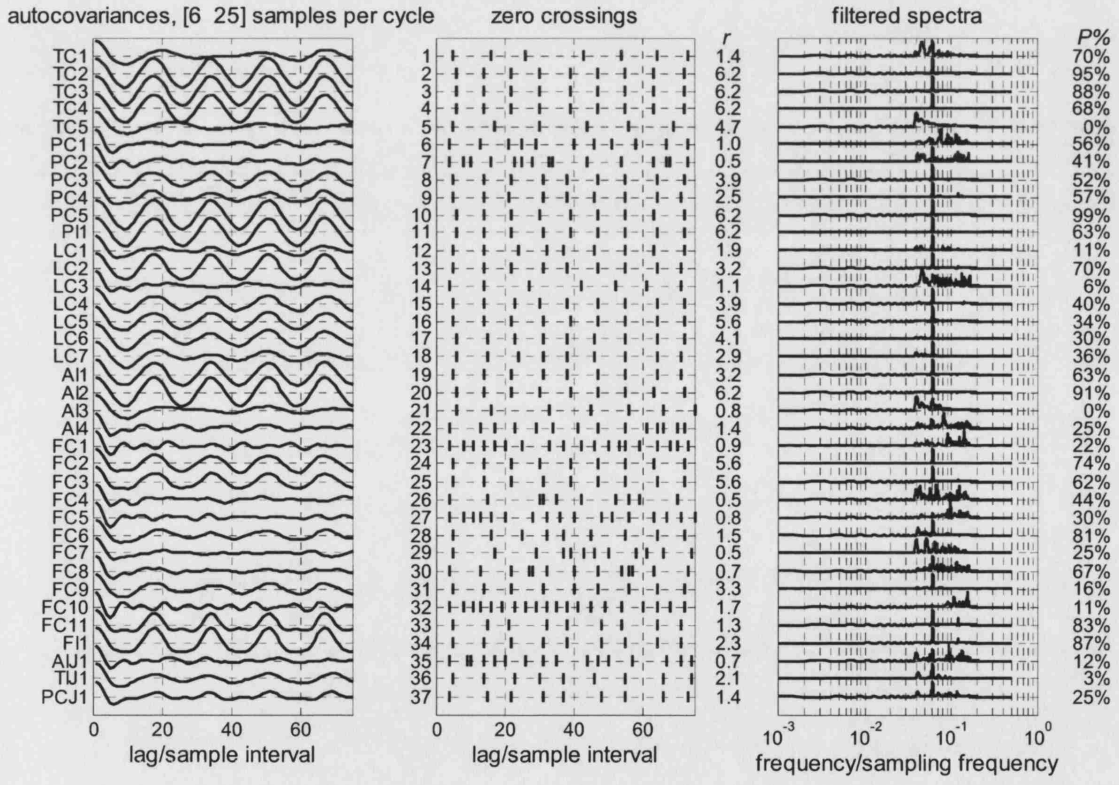

Fig 4.8 Plant-wide oscillation detection for SE Asia data in the range 6 to 25 samples per cycle. 23 tags are oscillating at about 16 samples per cycle (0.06 on the frequency axis) with oscillation index above 1 and power above 10%.

4.5 CHAPTER SUMMARY

The chapter has described and demonstrated the detection of distributed oscillations in a chemical process. A zero crossing technique was used which was greatly enhanced by use of the autocovariance functions (*ACF*) of the time trends. A statistic was devised to test the null hypothesis of random zero crossings having the exponential distribution. The test is sensitive to the regularity of the zero crossing on the lags axis and thus detects if an oscillation is present. If so, its period is determined from the intervals between the zero crossings.

The *ACF* of an oscillatory time trend is also oscillatory and the zero crossings are much easier to detect accurately because the *ACF* not noisy. White noise in a time trend all maps to the zero lag channel in the *ACF* and does not disturb the *ACF* at larger lags. However, other oscillations can cause interference at larger lags and will upset the regularity of the zero crossings. Frequency domain filtering was used to address this issue with the filters being designed by an algorithm to isolate each oscillation in turn.

A further contribution of the work has been an automated method for grouping the oscillations. The criterion for membership of a group is based on the period of oscillation of the tag under test T_p and the observed variation in its period, σ_{T_p}. The condition is:

$$\left|T_p - T_0\right| < \sigma_{T_p}$$

where T_0 is the period of the most regular oscillation in the group. It means the period of oscillation should be no more different than one standard deviation from the period of the most regular oscillation.

The final outcome of an oscillation analysis is a report stating the groups of oscillations detected and the tags participating in that oscillation. The chapter thus achieves the aim of detection of distributed oscillations. However, not all distributed disturbances are oscillatory. The next chapter will extend the work to the detection of distributed disturbances with arbitrary non-oscillating forms.

5. SPECTRAL PRINCIPAL COMPONENT ANALYSIS

5.1 OVERVIEW

5.1.1 Motivation

Distributed disturbances in chemical processes can be well characterized by their power spectra. In some cases they may be oscillatory and have sharp spectral peaks while in other cases they may be more complicated having broad band features or multiple overlapping spectral peaks. A method is therefore needed for detection and characterization of these more general types of disturbances. Topics of interest include:

- Detection of the presence of plant-wide disturbances indicated by a specific pattern of spectral features;
- Characterization of those spectral features;
- Determination of the groups of tags affected by the disturbances

The method must be applicable to very large data sets because it will provide a broad-brush analysis at the initial stage of a detection and diagnosis study in order to identify all plant disturbances. There is thus a requirement for a means of visualization of the results from large data sets with multiple types of behaviour present.

5.1.2 Introduction

Chapter 5 of the thesis describes the methods of spectral principal component analysis (PCA) that have been created to aid the initial stages of a detection and diagnosis study. They provide a classification of all the disturbances present and identification of the tags that are participating in those disturbances. The concepts are illustrated with data from a hydrogen plant in a SE Asian refinery that was introduced in Chapter 4.

The next section gives the background to spectral PCA and explains the advances provided by the work of the thesis. It also gives the formulation and methods for spectral PCA. Section 5.3. gives the results of a spectral PCA analysis of the SE Asia case study. The study indicates that the conventional methods for visualization of results from a principal component analysis by

means of score and loading plots is not adequate for a large-scale example since the SE Asia data needed seven principal components and hence a seven dimensional space to visualize the spectral relationships between the tags. Section 5.4. provides a new way to visualize the results of a high-dimensional principal component analysis by means of a hierarchical classification tree. The final section gives a summary of the findings of disturbance detection in the SE Asia data. It compares the findings from oscillations and spectral PCA, locates the disturbed tags on the process schematic and gives some preliminary observations about possible root causes.

5.1.3 Work done by the thesis author and others

The formulation of spectral PCA and the SE Asia case study reported in this chapter have been published in the following paper:

> Thornhill, N.F., Shah, S.L., Huang, B., and Vishnubhotla, A., 2002, Spectral principal component analysis of dynamic process data, *Control Engineering Practice*, **10**, 833-846.

The co-authors are Professors Sirish Shah and Biao Huang of the University of Alberta and Anand Visnubhotla of Matrikon Inc, Edmonton, Canada. The paper was written by Thornhill who also created all software, results, diagrams and graphs and the co-authors read, discussed and criticized the text and helped to give an industrial relevance.

5.2 DETECTION OF DISTURBANCES USING SPECTRAL PCA

5.2.1 Introduction to spectral principal component analysis

Overview: Principal component analysis (PCA) is a method of multivariate analysis. Chapter 2 introduced the methods and gave the key literature references. As formulated in this chapter the purpose of PCA is to discover a set of basis functions from which all the measurements in the plant can be approximately reconstructed. It is well known, however, that plant measurements having time delays or phase lags caused by process dynamics cause difficulties for principal component analysis in the time domain because dynamic features become misaligned. It is therefore not possible for time domain PCA to recognize them as similar unless some special procedure such as time shifting is applied to align features. For instance, time trends oscillating

as $\cos(\omega t)$ and $\sin(\omega t)$ are identical apart from a phase lag of $\pi/2$ rad, but since cos and sin are orthogonal functions a time-domain PCA does not recognize their similarity.

Spectral principal component analysis applies PCA to spectra rather than to the time trends and has several advantages over analysis in the time domain. Power spectra are invariant to time delays and therefore spectral PCA bypasses the need for time shifting to align dynamic features as shown by the observation that the power spectra of $\cos(\omega t)$ and $\sin(\omega t)$ are identical. Spectral PCA gives improved signal to noise ratio if the spectral content of the wanted signal occupies a narrow band of frequencies compared to the noise. That is to say, a narrow band spectral peak stands out from broad-band noise even when the r.m.s. value of the time-domain signal is smaller than that of the contaminating noise. The power spectrum is also insensitive to missing values and outliers because the transforms of such effects are spread thinly across all frequencies in the spectrum, whereas missing values in time-domain PCA can be quite problematical.

Contributions of this thesis: The work reported in this chapter overcomes the problem of alignment of time trends in a PCA by the use transforms of the time series that are invariant to time delay. The journal paper based on this work (Thornhill *et. al.*, 2002) is the first report in the literature of the application of PCA to the spectra of process measurements. It provides a means for historical auditing of data sets and is especially good for detection of groups of measurements with the same dynamic features in their time trends and characterization of those features in terms of the spectral content.

A second major extension is the visualization of the results of spectral PCA analysis using n principal components for which an $n-$dimensional score plot is needed to see the relationships between measurements. A method has been devised to show the relationships on a hierarchical classification tree of the type described originally by Chatfield and Collins (1980). This advance enables rapid and quantitative identification of tags whose spectra have similar features in a data set where large numbers of disturbances are present.

5.2.2 Techniques of spectral principal component analysis

Formulation of spectral PCA: In spectral PCA the rows of the data matrix, $\boldsymbol{X}$, are the power spectra $P(f)$ of the signals over a range of frequencies up to the Nyquist frequency (one half the sampling frequency). It is only necessary to use frequencies up to one half of the sampling frequency because the spectral powers in frequency channels $k=0$ (the d.c. or zero frequency) up to $k=N/2$ are distinct while spectral powers at frequencies $k=(N/2)+1$ to $N-1$ are the same as the powers for channels $k=N/2-1$ back down to $k=1$. There are thus $(N/2)+1$ frequency channels in use, where N is the number of samples in the data set (N is assumed to be an even number).

$$N/2+1 \textit{ frequency channels} \rightarrow$$

$$\boldsymbol{X} = \begin{pmatrix} P_1(f_0) & \cdots & P_1(f_{N/2}) \\ .. & .. & .. \\ P_m(f_0) & \cdots & P_m(f_{N/2}) \end{pmatrix} \begin{matrix} m \textit{ process} \\ \textit{variables} \\ \downarrow \end{matrix}$$

A full PCA decomposition reconstructs the $\boldsymbol{X}$ matrix as a sum over m orthonormal basis functions $\boldsymbol{w}'_1$ to $\boldsymbol{w}'_m$ which are spectrum-like functions each having $(N/2)+1$ frequency channels arranged as a row vector:

$$\boldsymbol{X} = \begin{pmatrix} t_{1,1} \\ \cdots \\ t_{m,1} \end{pmatrix} \boldsymbol{w}'_1 + \begin{pmatrix} t_{1,2} \\ \cdots \\ t_{m,2} \end{pmatrix} \boldsymbol{w}'_2 + \ldots + \begin{pmatrix} t_{1,m} \\ \cdots \\ t_{m,m} \end{pmatrix} \boldsymbol{w}'_m$$

The above expression may be written compactly as $\boldsymbol{X} = \boldsymbol{TW}'$ where the i'th column of $\boldsymbol{T}$ is $\begin{pmatrix} t_{1,i} & \cdots & t_{m,i} \end{pmatrix}'$ and the rows of $\boldsymbol{W}'$ are $\boldsymbol{w}'_1$ to $\boldsymbol{w}'_m$. The orthonormality of the rows of $\boldsymbol{W}'$ means that $\boldsymbol{T} = \boldsymbol{XW}$. Singular value decomposition $\boldsymbol{X} = \boldsymbol{UDV}'$ provides a means for computation of the requisite vectors with $\boldsymbol{T} = \boldsymbol{UD}$ and $\boldsymbol{W}' = \boldsymbol{V}'$. Matrix $\boldsymbol{D}$ of dimension $m \times m$, it is diagonal and its elements are the positive square roots of the eigenvalues of $\boldsymbol{XX}'$. Matrix $\boldsymbol{T}$ is $m \times m$ and $\boldsymbol{W}'$ has m rows and $(N/2)+1$ columns.

Choosing the number of PCs: A description of the majority of the variation in $\boldsymbol{X}$ can often be achieved by truncating the PCA description. If all the variables had similar spectral features then one term (i.e. one principal component) would describe most or all of the variability in the spectra. In other cases more terms may be needed. The following is a three-PC model in which the variation of $\boldsymbol{X}$ that is not captured by the first three principal components appears in an error matrix $\boldsymbol{E}$:

$$\boldsymbol{X} = \begin{pmatrix} t_{1,1} \\ \cdots \\ t_{m,1} \end{pmatrix} \boldsymbol{w}'_1 + \begin{pmatrix} t_{1,2} \\ \cdots \\ t_{m,2} \end{pmatrix} \boldsymbol{w}'_2 + \begin{pmatrix} t_{1,3} \\ \cdots \\ t_{m,3} \end{pmatrix} \boldsymbol{w}'_3 + \boldsymbol{E}$$

The issue of the correct number of terms is discussed by Chatfield and Collins (1980), Valle *et.al.*, (1999) and elsewhere. It is often decided on the basis of the eigenvalues λ_i of the data covariance matrix $\boldsymbol{XX}'$ which represent the variance in the data captured by the corresponding $\boldsymbol{w}'_i$ – vectors. For instance, the SE Asia data captured 90% of the variability in the data set in three PCs meaning that $(\lambda_1 + \lambda_2 + \lambda_3)/\sum \lambda_i \geq 0.9$ i.e. that the sum of the first three eigenvalues came to more than 90% of the total.

Graphical representation: Each spectrum in $\boldsymbol{X}$ may be represented graphically. For instance, when three $\boldsymbol{w}'$ – vectors are in use the i'th spectrum maps to a point having the co-ordinates $t_{i,1}$,

$t_{i,2}$ and $t_{i,3}$ in a three-dimensional space called a scores plot. Similar spectra have similar *t*-coordinates. Therefore such groups form clusters.

The rows of the error matrix, $\boldsymbol{E}$, are the unmodelled part of the spectra in $\boldsymbol{X}$. Therefore the magnitudes of the rows of $\boldsymbol{E}$ should be inspected. The following vector of squared prediction errors (SPE) can be used in which each term is the sum of squares of the elements of a row of $\boldsymbol{E}$ (Jackson and Mudholkar, 1979; Wise *et.al.* 1990; Kourti and McGregor, 1996).

$$\boldsymbol{SPE} = \begin{pmatrix} SPE_1 \\ SPE_2 \\ \cdots \\ SPE_m \end{pmatrix} = \begin{pmatrix} \sum_{k=0}^{N/2} (E_{1,i})^2 \\ \sum_{k=0}^{N/2} (E_{2,i})^2 \\ \cdots \\ \sum_{k=0}^{N/2} (E_{m,1})^2 \end{pmatrix}$$

A common reason why one particular spectrum may have a large SPE_i value (i.e. it is not well modelled by a reduced number of principal components) is that it has spectral features at frequencies where no other spectrum shows any features. The distribution is plotted as a histogram of SPE_i values and abnormal tags are taken to be those that were judged to be outliers in the histogram

Data pre-processing and filtering of spectra: It is usual in time-domain PCA to mean centre the data sets and to scale to unit variance. In spectral PCA the following data pre-processing may be applied:

- Mean centre the time trend and remove linear trends before calculation of the spectra;
- Filter the spectra, if required;
- Scale the spectra to the same total power such that:

$$\sum_{k=0}^{N/2} P(f_i) = constant$$

where $P(f_i)$ is the spectral power in the i'th channel;

The first of these steps removes steady offsets or linear ramps from the time domain data which would otherwise appear in the $k = 0$ channel of the spectrum.

The spectrum may be filtered by replacing the values in selected frequency channels with zeros, as described in Chapter 3. The ease of such an approach is a benefit of spectral PCA in contrast to filtering in the time domain where design of a narrow-band filter is a skilled task.

The scaling steps are equivalent to scaling to constant variance in the time domain because the signal variance is proportional to total power. Alternatively the spectra may be left unscaled if a meaning can be attributed to the intensity of the signal as is the case with acoustic monitoring, for instance.

5.2.3 PCA using autocovariance functions and time shifting

Supplementary methods: This sub-section briefly describes an extension to spectral PCA using autocovariance functions rather than power spectra. It has benefits for very large data sets. Also presented here is a method for time shift PCA. The reason for presenting it is so that the benefits of the new spectral PCA method can be compared with the standard traditional method.

Autocovariance PCA: When the autocovariance is determined from the inverse Fourier transform of the two-sided power spectrum then the advantages of use of spectra highlighted in the introduction apply also to autocovariance functions. In particular, like power spectra, autocovariance functions are invariant to time lags and delays. For example, $\sin(\omega t)$ and $\cos(\omega t)$ have the same autocovariance function. Principal component analysis using autocovariance covariance functions is formulated in the same way as was described for spectral PCA except that the rows of the data matrix are the autocovariance functions for the time trends. In that case, the $\boldsymbol{w}'-$vectors are an orthonormal set of basis functions for the autocovariance functions. The preprocessing steps are:

- Mean centre the time trend and remove linear trends before calculation of the spectra;
- Filter the spectra, if required;
- Create the autocovariance functions as the inverse DFT of the two-sided power spectra;
- Scale the time trend to unit standard deviation so that the covariance at zero lags is unity. For a zero mean scaled signal the autocovariance function is the same as the autocorrelation function.

In PCA using autocovariance functions $(N/2)+1$ lags from 0 up to $N/2$ may be used where N is the number of samples in the original data set. The performance of PCA using *ACF*s has been shown to be identical to spectral PCA using $(N/2)+1$ frequency channels (Thornhill *et. al.* 2002). The reason for this is that the information encoded in the *ACF* is a linear transform of the information in the spectrum with $(N/2)+1$ frequency channels.

Dealing with large data sets: If the data sets are so large that computing speed or memory is an issue it may become necessary to reduce the size of the data matrix to make the problem computable. One choice is lower resolution single-sided power spectra calculated from the averaged periodogram method (Welch, 1967). In that case the number of frequency channels, M, may be considerably fewer than the original number of data points and data compression is achieved. An alternative is to calculate the autocovariance function from the whole time trend and then to use just its first M lags in the PCA calculations.

The information contents of a Welsh spectrum with M channels and the autocovariance function with M lags are *not* the same. Use of the M channel autocovariance function has been shown to have superior performance compared to the use of a low resolution spectrum with M frequency channels when $M \ll (N/2)+1$ (Tan *et. al.*, 2002). The reason is that the user selects M so that

the M autocovariance lags capture several cycles of the features of interest. These are often at mid-range or high frequency. In effect, by choosing a shortened autcovariance function the user focuses on the high frequencies and ignores low frequencies whereas reduction of the data by means of the Welsh spectrum degrades the resolution right across the whole frequency range. In cases where size reduction is needed and when both low and high frequencies are of interest the Welsh spectrum with M channels would be expected to be better. An ACF truncated to M lags does not capture low frequencies adequately because the longest complete cycle contained in M lags has a frequency of $1/M$.

Time-shift PCA: The method of lagged variables has been used in time-domain PCA to address the problems of time lags and delays. In this thesis the results of time-shift PCA are presented for comparison to show that spectral PCA has better performance. Wise and Gallagher (1996), gave an overview of the procedure. If the time delays are known then time trends in the rows of the $\boldsymbol{X}$ matrix may be replaced by time delayed trends:

$$\begin{matrix} & \textit{n samples} \rightarrow & \\ \boldsymbol{X} = \begin{pmatrix} x_1(t_1) & \cdots & x_1(t_n) \\ x_2(t_1-\tau_2) & \cdots & x_2(t_n-\tau_2) \\ \cdots & \cdots & \cdots \\ x_m(t_1-\tau_m) & \cdots & x_m(t_n-\tau_m) \end{pmatrix} & \begin{matrix} \textit{m process} \\ \textit{variables} \\ \downarrow \end{matrix} \end{matrix}$$

where the row vectors x_i are the time trends of m measurements having n samples and τ_2 to τ_m are the time shifts of the second to m'th measurements. The time shifts are relative and can be negative. For instance, τ_i would be negative if a feature in the i'th time trend occurred before a similar feature in the reference time trend in the first row. In the case of a single plant-wide oscillation, alignment of the measurements using the correct time delays would enable a one-PC model to describe most or all of the variability. The $\boldsymbol{w}'$ – vector would be a periodic time domain function with the same frequency as the oscillating measurements.

Determination of the time shifts: The arguments of the Fourier transforms of the time trends can be used to determine the optimum time shifts for the case when the time trends are oscillating with a single frequency. If the angular frequency of oscillation is ω_{osc} then the phase difference between the i'th process variable and the first is:

$$\alpha_i = \angle X_i(j\omega_{osc}) - \angle X_1(j\omega_{osc})$$

where $\angle X_i(j\omega_{osc})$ is the argument of the complex Fourier transform of $x_i(t)$ at the frequency of the oscillation (i.e. at the frequency where the power spectrum has its peak). The frequency response of a pure delay T_d is $e^{-j\omega T_d}$. This is a complex quantity having unity magnitude and argument $-\omega T_d$, meaning that the phase of the delayed signal relative to the undelayed signal is $-\omega T_d$. Therefore the time delay of the i'th process variable relative to the first is $T_{d_i} = -\alpha_i/\omega_{osc}$.

5.3 WORKED EXAMPLE IN SPECTRAL PCA

5.3.1 Basic spectral PCA

Overview: As with oscillation detection, spectral PCA analysis is done on the controller errors rather than on the *pv*'s because in some cases the controller set points are varying. The focus should be on disturbances, not on variability induced because of set point changes. In this section spectral PCA is applied to the SE Asian refinery data described in Chapter 4. The time trends of the controller errors and spectra are presented again in Figure 5.1.

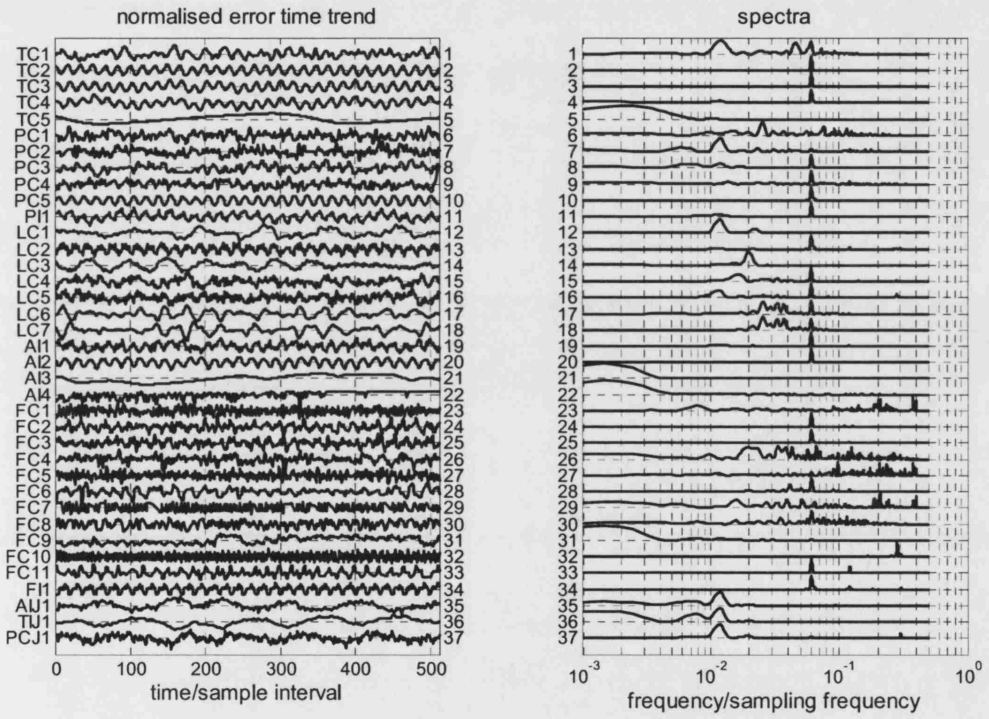

Fig 5.1 Controller errors and spectra for the SE Asia data set. The aim is to detect tags whose spectra are similar.

Spectral PCA reconstruction: The values presented in the middle column of Table 5.1 show the amount of variability captured. They are ratios expressed as percentages between the first seven eigenvalues of $\boldsymbol{XX}'$ and the sum of all the eigenvalues, where $\boldsymbol{X}$ is the data matrix containing the spectra. The cumulative results in the right hand column show that a three-PC model captured 90% of the spectral variability and that a seven-PC model captured almost 97% of the variability.

PC number	spectral variation captured	cumulative spectral variation captured
1	66.7%	66.7%
2	16.6%	83.3%
3	7.1%	90.4%
4	2.6%	93.0%
5	1.7%	94.7%
6	1.1%	95.8%
7	0.9%	96.7%

Table 5.1 Spectral variability captured by spectral PCA.

A three-PC model: The reason for presenting a three-PC plot is that it is common practice in the literature to present two– or three–PC score plots. Figure 5.2 shows the loading and SPE plots. It is clear from the SPE plot that a three–PC model has failed to adequately model the spectra of tags 5, 14, 17, 18 and 32 because these all have large SPE values. A seven–PC model will be presented in section 5.4 that has better performance with these tags.

The three–PC score plot is shown in Figure 5.3 where the upper panel shows the major groups of tags having large score values and the lower panel shows a close-up view of tags with scores near the origin. These figures will be discussed in more detail in the next sub-section.

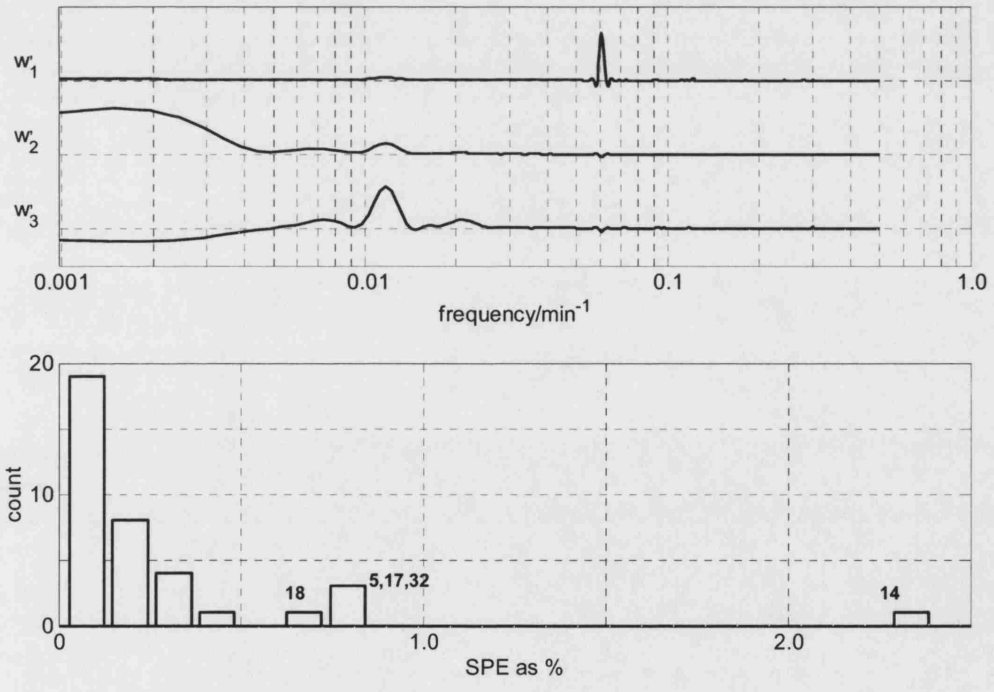

Fig 5.2 A three-PC model for the SE Asia data. Upper panel: Loading vectors. Lower panel: SPE values showing the numbers of tags having *SPE*s in various ranges. Tag 14 has the largest *SPE*.

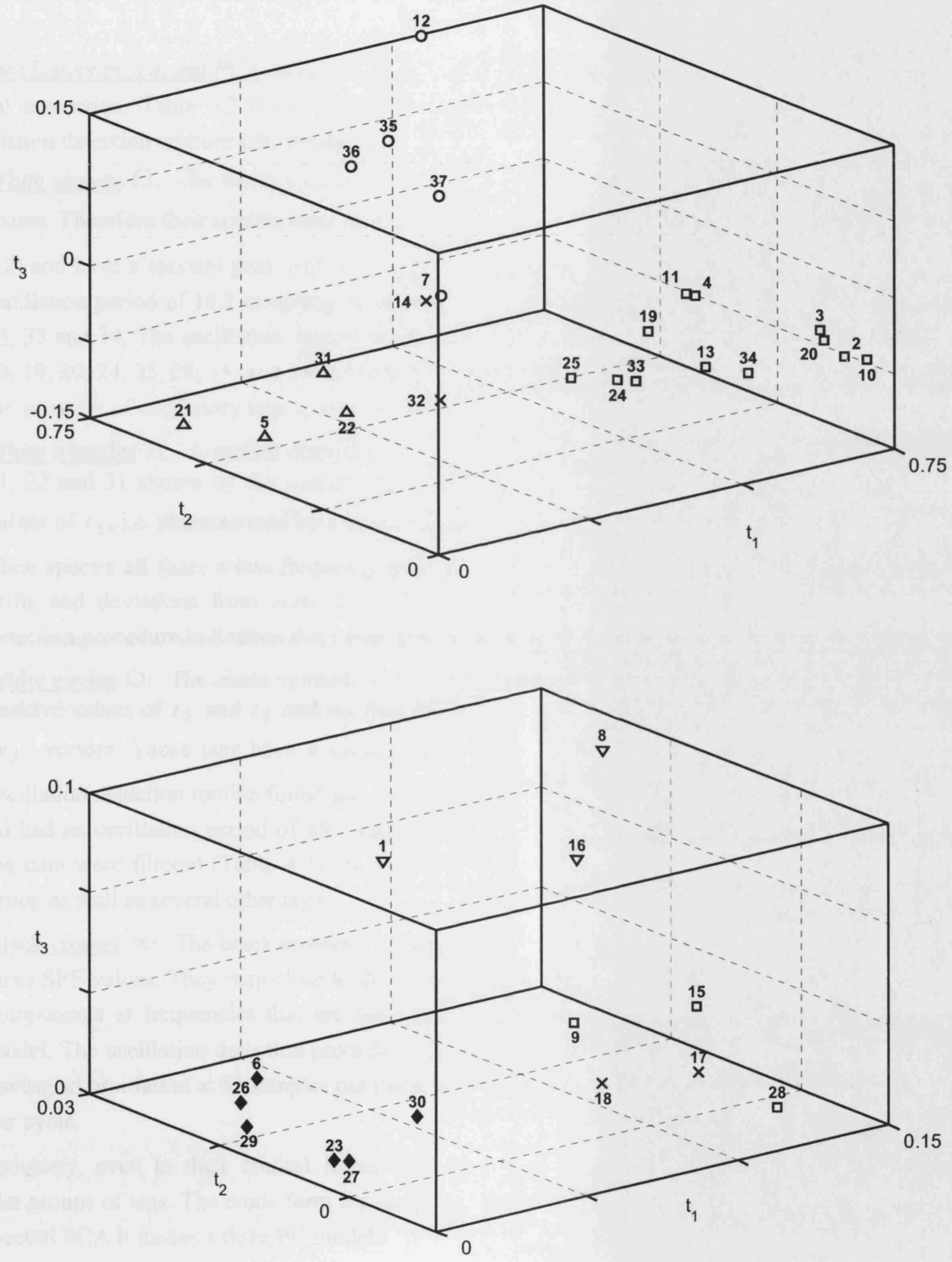

Fig 5.3 Three-PC score plots with tag numbers for SE Asia data. Upper panel: Major clusters forming plumes radiating from the origin. Lower panel: Close up of the score plot near the origin.

5.3.2 Discussion of the PCA clusters

Major clusters in spectral PCA: Major clusters in the spectral PCA score plot can be identified by visual inspection. Table 5.2 below compares the findings from a three–PC model and basic oscillation detection without filtering from section 4.4.1.

White squares □: The white squares in figure 5.3 represent a family of tags with positive t_1 scores. Therefore their spectra were dominated by the w'_1 – loading (see upper panel of Figure 5.2) and have a spectral peak with a frequency of 0.06 on the normalised frequency axis. (an oscillation period of 16.7 sampling intervals). These were tags 2, 3, 4, 10, 11, 13, 19, 20, 24, 25, 33 and 34. The oscillation detection results in section 4.4.1 indicated 2, 3, 4, 8, 9, 10, 11, 13, 19, 20, 24, 25, 28, 33, and 34 and therefore the three–PC spectral PCA model has detected the same set of oscillatory tags apart from 8, 9 and 28.

White triangles △: A second distinct cluster of tags is the white triangles. These were tags 5, 21, 22 and 31 shown by the spectral PCA plot to have positive values of t_2 and negative values of t_3, i.e. characterized by a weighted difference between the w'_2 – and w'_3 – vectors. Their spectra all share a low frequency feature which means the time trends show long term drifts and deviations from zero. This group of tags was not detected by the oscillation detection procedure in Section 4.4.1 because these tags are not oscillatory.

White circles ○: The circle symbols in Figure 5.2 are tags 7, 12, 35, 36 and 37 which have positive values of t_2 and t_3 and are thus characterized by a weighted sum of the w'_2 – and w'_3 – vectors. Those tags have a spectral peak just above 0.01 on the frequency axis. The oscillation detection routine found some members of this group. In Table 4.4.1 tags 12, 35 and 36 had an oscillation period of 85.7 samples per cycle (0.012 on the frequency axis). When the data were filtered (Table 4.7) the oscillation detection procedure added 7 and 36 to the group as well as several other tags.

Black crosses ×: The black crosses in Figure 4.6 are for tags 14 and 32, both of which had large SPE values. They map close to the origin of the score plot because they have frequency components at frequencies that are not captured by the three w' – vectors of the three-PC model. The oscillation detection procedure also found that these were unique tags with tag 14 having an oscillation at 51 samples per cycle and 32 having a fast oscillation with 3.6. samples per cycle.

In summary, even in their crudest forms basic oscillation detection and spectra PCA found similar groups of tags. The crude form for oscillation detection means no filtering, and in the case of spectral PCA it means a three PC model.

Basic oscillation detection		**Three–PC spectral PCA**		
ave_period	tags	frequency	period	tags
3.56	32	–	–	–
16.4	2, 3, 4, 8, 9, 10, 11, 13, 19, 20, 24, 25, 28, 33, 34	0.06	16.7	2, 3, 4, 10, 11, 13, 19, 20, 24, 25, 33, 34
50.9	14	–	–	–
85.7	35 12 36	0.012	83.3	7, 12, 35, 36, 37
–	–	low	–	5, 21, 22, 31

Table 5.2 Comparison of basic oscillation detection and three–PC spectral PCA.

Minor clusters in spectral PCA: The lower panel of Figure 5.2. shows the tags whose spectra map close to the origin of the score plot.

Black diamonds ◆: The cluster near the co-ordinates (0,0,0) comprises tags 6, 23, 26, 27, 29 and 30. Their t_1, t_2 and t_3 scores were all close to zero because the magnitudes of their spectra were small, and the reason for that is that they have spectral features spread uniformly across a wide range of frequencies. The broad-band spectrum indicates that these tags comprised mainly random noise. Tags 23, 26, 27, 29 and 30 are flow loops while 6 is a pressure control loop. They have been tuned close to minimum variance and have a random error signal with no predictable features such as oscillations or steady offsets (Harris, 1989). The conclusion is that well tuned control loops map to the origin of the spectral PCA score plot. The conclusion can be inverted to give a strong guideline for process control engineers:

> Tags that do not map to the origin of the spectral PCA score plot have some undesirable behaviour that requires attention.

White squares □: The figure suggests that tags 9, 15, 17, 18, and 28 form a cluster in the lower right corner, but that impression is misleading because tags 17 and 18 have large *SPE* values and therefore are projections into the three-dimensional space from points in a higher PCA dimension. Therefore the proposed cluster would actually be only tags 9, 15 and 28. They have positive t_1 values showing they are characterized by the 0.06 min^{-1} spectral peak (16.7 sample per cycle).

Black crosses ×: Tags 17 and 18 having large *SPE* values are not well modelled by a three–PC model.

White inverted triangles ▽: The remaining tags are 1, 8 and 16. They are quite close to the origin because their time trends were noisy, but have positive values of t_1 and t_3 and thus feature the spectral peaks at around 0.01 min^{-1} and the 0.06 min^{-1} oscillation.

Plumes in the score plot: Figure 5.3 shows that similar spectra in spectral PCA form plumes of tags radiating from the origin of the score plot. The physical interpretation is simple in the case of the white squares in the upper panel of Figure 5.3, where the plume extends along the t_1 – axis of

the score plot and the t_2 and t_3 values are close to zero. Remembering that the three-PC model is:

$$X = \begin{pmatrix} t_{1,1} \\ \dots \\ t_{m,1} \end{pmatrix} w'_1 + \begin{pmatrix} t_{1,2} \\ \dots \\ t_{m,2} \end{pmatrix} w'_2 + \begin{pmatrix} t_{1,3} \\ \dots \\ t_{m,3} \end{pmatrix} w'_3 + E$$

then for the tags in the plume along the t_1 – axis the model is:

$$X_k \approx t_{k,1} \times w'_1 + 0 \times w'_2 + 0 \times w'_3 + E_k$$

where X_k is the k'th row of X and E_k is the k'th row of E. The loading vector w'_1 is shown in Figure 5.2 to have a strong spectral feature at 0.06 on the frequency axis (i.e. the 16.7 samples per cycle oscillation). The value of $t_{k,1}$ thus indicates how strong that oscillation is compared to the background noise in E_k. Tags with high spectral energy in that band have large $t_{k,1}$ values and are far from the origin, those where the oscillation is smaller and more swamped by noise have smaller $t_{k,1}$ values. Thus the oscillatory tags tend to line up along the t_1 – axis in the score plot in order of the strength of the oscillation. In reality most tags also have small non-zero scores on the t_2 – and t_3 – axes, but the overall effect is that the tags with the 16.7 samples per cycle oscillation form a plume radiating out in the direction of the t_1 – axis.

Similar comments apply to other plumes, however the plumes may not be exactly along an axis. The plume containing tags 7, 12, 35, 36 and 37 is not along an axis of the score plot. Instead, it is at a constant angle in the t_2 – and t_3 – plane, thus showing that the model for those tags is:

$$X_k \approx 0 \times w'_1 + t_{k,2} \times w'_2 + t_{k,3} \times w'_3 + E_k$$

with the scores $t_{k,2}$ and $t_{k,3}$ in a constant ratio to each other.

5.3.3 Comparison with time domain PCA

Three–PC time domain model: Table 5.3 shows that the use of PCA directly on the time trends (time domain PCA) in the refinery data set performed poorly. The first three terms in the model described only 37% of the variability in the data set and even seven PCs captured only about 57%. Figure 5.4 shows the scores plot of the time-domain PCA analysis and the w' – vectors, which are basis functions for reconstruction of the time trends. Clusters such as the white squares that were clear in spectral PCA analysis do not emerge in the time domain PCA analysis.

The white squares in Figure 5.4 are the tags oscillating at about 16.7 samples per cycle. They have large magnitudes of t_1 or of t_3, or of both and are therefore characterized by various weighted combinations of the w'_1 – and w'_3 – vectors and are oscillatory time domain functions. Vector w'_3 has a phase shift of one quarter of a cycle compares to w'_1. The explanation for the lack of clustering of the square symbols is that both w'_1 and w'_3 are needed in various amounts to

reconstruct oscillations of arbitrary phase. In spectral PCA, by contrast, the square symbols were clustered together because the power spectra are invariant to the phase of the oscillation. These results demonstrate that spectral PCA gives better performance than time domain PCA when process dynamics and time delays cause phase shifts in the time trends.

PC number	variation captured	cumulative variation captured
1	19.7%	19.7%
2	9.2%	28.9%
3	7.8%	36.8%
4	6.0%	42.8%
5	5.2%	47.9%
6	4.6%	52.5%
7	4.3%	56.8%

Table 5.3 Variability in the time-domain data captured by time-domain PCA.

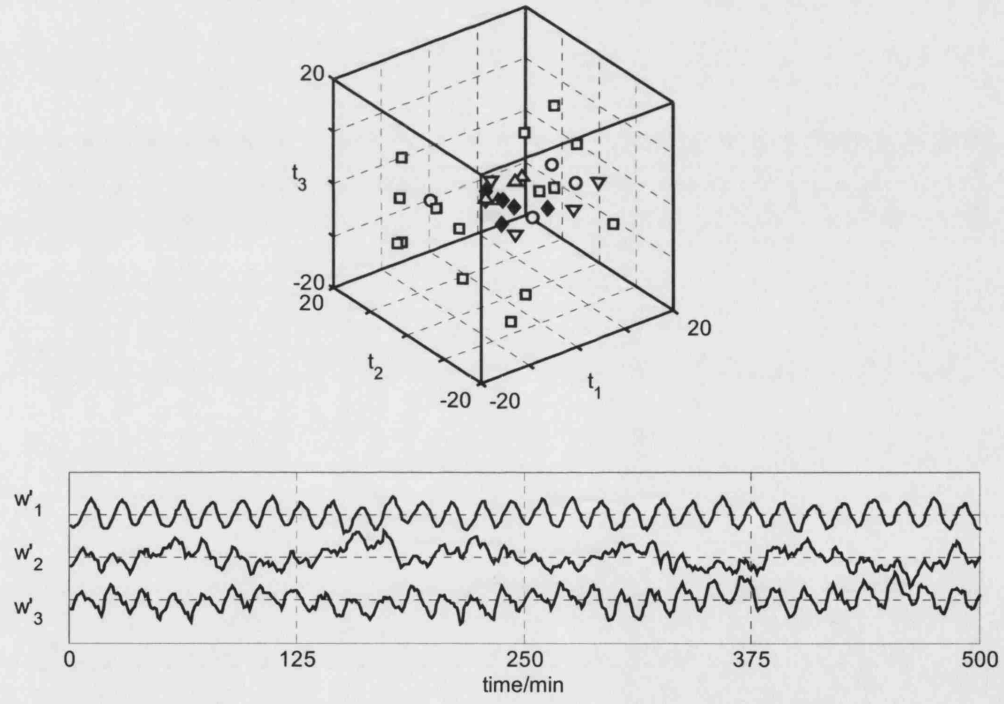

Fig 5.4. Three-PC score plots for time-domain PCA. The presence of time delays and phase lags destroys the clustering of similar tags.

Time shifted PCA: A further study was conducted using time shifted PCA on a subset of the refinery data. The tags used were 2, 3, 4, 10, 20 and 34 selected because they showed the clearest oscillation and their time shifts could be easily determined. The left hand plot of Figure 5.5 shows the selected time trends where it can be seen that the oscillations were not in phase and did not align. It also marks with a dot the time delays determined by the arguments of the Fourier

transforms at the frequency of oscillation using the expression derived earlier for the time delay of the *i*'th process variable relative to the *j*'th:

$$T_{d_{ij}} = -\frac{\angle X_i(j\omega_{osc}) - \angle X_j(j\omega_{osc})}{\omega_{osc}}$$

where $\angle X_i(j\omega_{osc})$ is the argument of the complex Fourier transform of $x_i(t)$ at the frequency of the oscillation (i.e. at the frequency where the power spectrum has its peak).

For instance, the arguments of the Fourier transform of tags 2 and 3 at $0.06\,\text{min}^{-1}$ were $0.165\,\text{rad}$ and $-1.043\,\text{rad}$ respectively. Therefore the phase of tag 3 relative to tag 2 was $-1.208\,\text{rad}$ and the time delay of tag 3 relative to tag 2 was:

$$T_d = \frac{1.208}{2\pi \times 0.06} = 3.204\,\text{min}$$

The right hand plot shows the time trends after time shifting in which it can be seen that the delays determined from the Fourier transforms have aligned the peaks and valleys.

Comparison of time domain, time shift and spectral PCA: A comparison was made of time domain PCA without time shifted vectors, PCA with time shifted vectors, and spectral PCA for these six tags. There is one dominant component of behaviour present (the 16.7 sample per cycle oscillation) and a second more minor component representing a higher frequency feature in Tag 34. Therefore most of the variability in this small data set can be captured by one PC while two PCs should capture all.

A one–PC spectral PCA model (98.9%) outperformed both time shifted PCA (85.7%) and time domain PCA (65.4%). Spectral PCA achieved almost 100% with a two–PC model, as expected, while time shifted PCA gave 93.2% and time domain PCA gave 90.4%.

It might be expected that the time-shifted PCA would have identical performance to spectral PCA because time shifting removes phase information. Therefore the finding that spectral PCA has superior performance needs an explanation. The reason is that the phase information was not fully removed by time shifting because time shifts can only be applied in integer multiples of the sampling interval. Hence, for instance, the time delay determined for tag 3 relative to tag 2 was 3.20 min, but the time shift applied was only 3 min being the closest integer number of sampling intervals.

In general, time shifting cannot remove all delays. For instance, suppose that oscillations at two unrelated frequencies originate in control loops A and B and propagate plant-wide. Propagation delays from A are not the same as from B because the routes through the plant are different and therefore it is not normally feasible to find a unique time shift that aligns both oscillations simultaneously.

It is concluded that spectral PCA gives significant advantages over time-shifted PCA for dynamic data. It has also been demonstrated that that time-domain PCA is not suitable for the analysis of dynamic data.

PC	time domain PCA	time shift PCA	spectral PCA
1	65.4%	85.7%	98.9%
2	90.4%	93.2%	99.7%
3	96.8%	97.5%	100.0%

Table 5.4 Cumulative variation captured by different PCA methods for the data set of Fig 5.5.

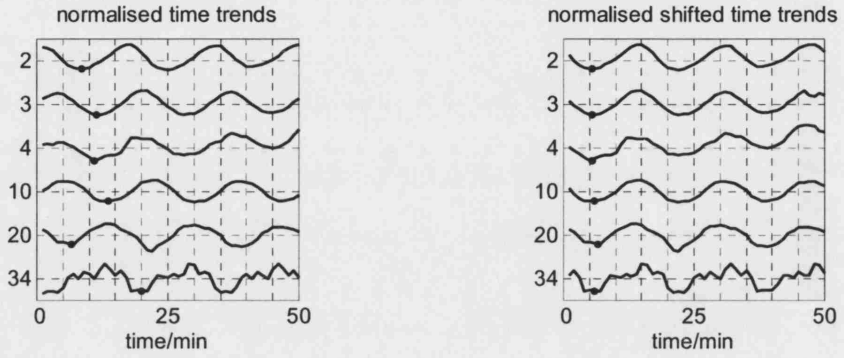

Fig 5.5 Oscillating time trends in the SE Asia data set: Left panel: Original data. Right panel: Time shifted to align the peaks.

5.4 HIGH DIMENSION SPECTRAL PCA MODELS

5.4.1 Score plots in higher dimensions

SPE values in a seven–PC model: In the three–PC model, Figure 5.2 showed tags 5, 14, 17, 18 and 32 had large *SPE* values which means their spectra have frequency components at frequencies that were not captured by the three $\boldsymbol{w}'$ – vectors of the model. For instance, tag 32 had an oscillation at a frequency of just under 0.2 min^{-1}. Figure 5.6 shows the $\boldsymbol{w}'$ – vectors for a seven–PC model:

$$\boldsymbol{X} = \begin{pmatrix} t_{1,1} \\ \dots \\ t_{m,1} \end{pmatrix} \boldsymbol{w}'_1 + \begin{pmatrix} t_{1,2} \\ \dots \\ t_{m,2} \end{pmatrix} \boldsymbol{w}'_2 + \dots\dots + \begin{pmatrix} t_{1,7} \\ \dots \\ t_{m,7} \end{pmatrix} \boldsymbol{w}'_7 + \boldsymbol{E}$$

The oscillation at 0.2 on the frequency axis in Tag 32 is captured by the w'_7 – vector. Figure 5.7 shows the SPE plot of the seven-PC model. No tags have large SPE values showing that the seven–PC model has captured all the spectral features of interest in the data set.

The seven-PC score plot: There is an issue, however, of visualisation of the score plot for a seven–PC model because seven orthogonal axes are required to plot the co-ordinates $\{t_{i,1}\ t_{i,2}\ t_{i,3}\ .....\ t_{i,7}\}$ of the i'th tag. A solution to this problem is explored in this section of the thesis.

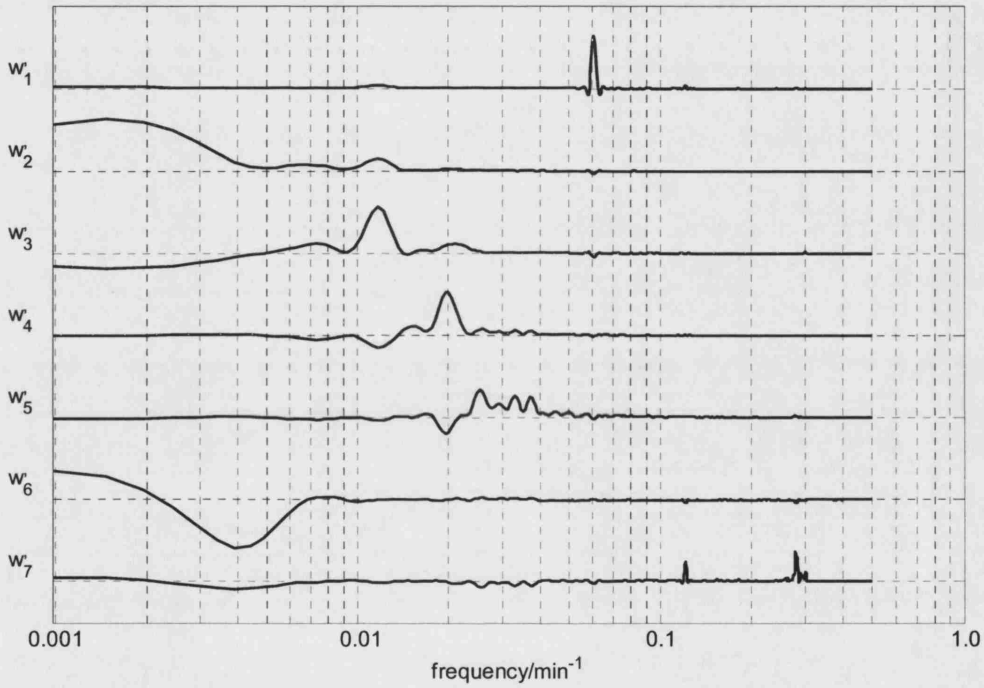

Fig 5.6 Loading vectors of a seven-PC model. The high frequency spectral features of tag 32 appear in w'_7, the tag 14 oscillation at 0.02 on the frequency axis is in w'_4.

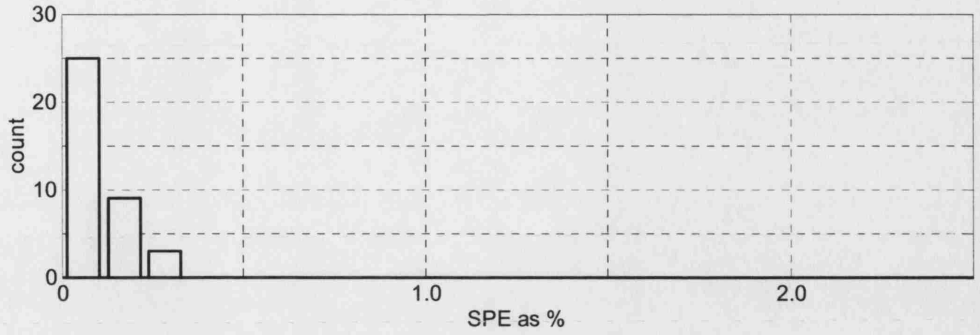

Fig 5.7 SPE values for a seven-PC model. None has a large SPE showing that $\boldsymbol{w}'_1$ to $\boldsymbol{w}'_7$ adequately capture almost all the spectral variability.

5.4.2 Visualization of score plots in higher dimensions

Distance measures for spectral PCA: Although it is only possible to visualize a maximum of three dimensions, the notion of a score plot can be generalized to higher dimensions. The approach explored here uses a hierarchical classification tree that shows which spectra in the seven-dimensional space are close to one another according to some suitable distance measure. Two distance measures are discussed, the Euclidian distance and an angle measure.

The Euclidian distances between scores in the score plot has been widely used as a measure for detection of PCA clusters. For instance, if seven principal components were used then the vector $\boldsymbol{s}_i$ joining the origin to the i'th spot in the score plot is of the form:

$$\boldsymbol{s}_i = t_{i,1}\hat{\boldsymbol{z}}_1 + t_{i,2}\hat{\boldsymbol{z}}_2 + t_{i,3}\hat{\boldsymbol{z}}_3 + t_{i,4}\hat{\boldsymbol{z}}_4 + t_{i,5}\hat{\boldsymbol{z}}_5 + t_{i,6}\hat{\boldsymbol{z}}_6 + t_{i,7}\hat{\boldsymbol{z}}_7$$

where $\hat{\boldsymbol{z}}_1$ to $\hat{\boldsymbol{z}}_7$ are orthogonal unit vectors. The Euclidian distance between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ is:

$$d_{i,j} = \sum_{k=1}^{7} \left(t_{i,k} - t_{j,k}\right)^2$$

In spectral PCA it is likely that an angle measure may be more suitable than the Euclidian distance. It was observed in Figure 5.3 that the tags with similar spectra in the score plot do not form tight clusters. Rather, they take the form of a plume radiating from the origin in a specific direction. An appropriate measure for membership of a plume is that direction of vector $\boldsymbol{s}_i$ in the score plot lies within the same small solid angle as those of other $\boldsymbol{s}-$vectors belonging to tags within the plume.

The angle between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ may be determined through calculation of the scalar product:

$$\cos\left(\theta_{i,j}\right) = \frac{\boldsymbol{s}_i \bullet \boldsymbol{s}_j}{\left|\boldsymbol{s}_i\right|\left|\boldsymbol{s}_j\right|}$$

where, for example, in the case of a seven PC model:

$$\boldsymbol{s}_i \bullet \boldsymbol{s}_j = \sum_{k=1}^{7} \left(t_{i,k} \times t_{j,k} \right) \text{ and } \left| \boldsymbol{s}_i \right| = \sqrt{\sum_{k=1}^{7} t_{i,k}^{\,2}}$$

It is noted that $\left\{ t_{i,1} \;\; t_{i,2} \;\; t_{i,3} \;\; \ldots\ldots \; t_{i,7} \right\}$, is the i'th row of matrix $\boldsymbol{T}$ in $\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}' + \boldsymbol{E}$ when a seven PC model is in use. Therefore the following calculation steps lead to a symmetric matrix whose elements are the cosines of the wanted angles $\theta_{i,j}$:

Algorithm: Calculation of the angle measures

Step 1: Create a normalized matrix $\tilde{\boldsymbol{T}}$ from $\boldsymbol{T}$ whose row vectors are of unit length. Each element in the i'th row of $\boldsymbol{T}$ is divided by $\sqrt{\sum_{k=1}^{N_{PC}} t_{i,k}^{\,2}}$ where N_{PC} is the number of principal components in use.

Step 2: Determine the angle matrix $\boldsymbol{\Omega} = \tilde{\boldsymbol{T}}\tilde{\boldsymbol{T}}'$. The elements of $\boldsymbol{\Omega}$ are $\cos\left(\theta_{i,j}\right)$. For instance, $\boldsymbol{\Omega}(15,17)$ is the cosine of the angle between $\boldsymbol{s}_{15}$ and $\boldsymbol{s}_{17}$, the vectors from the origin to the spots in the score plot representing the fifteenth and seventeenth spectra respectively.

Step 3: Create matrix $\boldsymbol{D}$ from:

$$\boldsymbol{D} = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & \\ \ldots & \ldots & & \\ 1 & & & \end{pmatrix} - \boldsymbol{\Omega}$$

It is convenient in practice to do the above transform in order to work with a distance matrix whose elements are $1 - \cos\left(\theta_{i,j}\right)$ because then distance measure is small when the spectra are close and larger when the angle between them is greater.

Automatic identification of spectral PCA plumes: The matrix $\boldsymbol{D}$, whose elements are $d_{i,j} = 1 - \cos\left(\theta_{i,j}\right)$, is to be analyzed to find high-dimensional plumes in the spectral PCA score plot. Two tags whose $\boldsymbol{s}-$vectors in the score plot point in similar directions have small values of $d_{i,j}$. An algorithm from Chatfield and Collins (1980) called hierarchical clustering is used to construct a hierarchical tree to aid the detection of plumes. The benefit of the hierarchical tree is that it is easily visualized even though it represents a high dimensional space.

Algorithm: Automated generation of PCA clusters

Step 1: The smallest non-zero value $d_{i,j} = d_1$ in the matrix is identified. Its row and column indexes i and j indicate which two spectra have the smallest angular separation and these are grouped as a cluster characterised by the distance d_1.

Step 2: A smaller matrix is then generated from the original. It does not have rows and columns for the two similar spectra identified at step 1. Instead, it has one row and column that give the distances of all the other spectra from the cluster. The distance is

$\min\{d_{i,n}, d_{j,n}\}$, i.e. the angular distance between the *n*'th spectrum and whichever member of the cluster was closer.

Step 3: The above procedure is repeated until all the spectra have been placed in clusters. At any stage, the outcome of the next step is either another spectrum added to a cluster already identified or the combining of two spectra to start a new cluster.

The graphical representation of the hierarchical tree based upon the clusters detected by the above algorithm has been programmed in the MATLAB programming language. It utilises a recursive algorithm for searching the clusters starting at the top and searching through sub-clusters and sub-sub-clusters to create the branches of the tree until the search terminates in a "leaf", meaning a single tag. The end result is a tree structure from which clusters can readily be visually identified such as the one shown in Figure 5.8.

5.4.3 Worked example of score plot visualization

Visualization of the seven-PC score plots using the angle measure: Figure 5.8 shows the classification of the spectra by angular separation in the seven-dimensional score plot. The numerical values indicate angular separation. For instance, tag 14 splits off at about 0.43 on the vertical axis. Thus $1-\cos(\theta_{i,14})=0.43$ and $\theta_{i,14}=55.2^{o}$, meaning that the angular separation between tag 14 in the seven-dimensional score plot and the next closest tag is 55.2^{o}.

Each node in the tree indicates that two sub-clusters have been identified. For instance, the node in the centre of the plot at 0.32 on the vertical axis shows a right-hand branch that terminates in a cluster for tags 5, 21, 22 and 31. The left hand branch divides and subdivides several more times eventually identifying sub-clusters such as 7, 12, 35, 36 and 37. The numerical value of $1-\cos(\theta_{i,j})=0.32$ means that the minimum angular separation between the {5, 21, 22, 31} group and any other tag was 47.2^{o}.

Twigs in the classification tree are the vertical lines attached to the individual tags and a *branch* is a vertical line in the tree which terminates in twigs. Tags are considered to form a cluster when the branch is longer than the twigs.

For example, Tags 5, 21, 22 and 31 form a cluster because the twigs are much shorter than the branch, and moreover 21, 22 and 31 are a sub-cluster with its own little branch. By contrast, tags 6, 26, 29 are not a cluster because the twigs are longer than the branch.

Results from a hierarchical tree using angle measures: Figure 5.8 shows groupings of tags that have small angular separation, i.e. groups of tags that form a plume. Three large clusters or plumes are present. The same symbols have been used in the three-PC score plot of Figure 5.2 and the seven-PC hierarchical classification tree of Figure 5.8.

- Tags with □ symbols: 2, 10, 3, 20, 24, 13, 34, 33, 25, 4, 11 and 19
- Tags with ○ symbols: 7, 12, 35, 36 and 37
- Tags with △ symbols: 5, 21, 22 and 31
- Tags 17 and 18 form an additional smaller sub-cluster as do 23 and 27

Results from a hierarchical tree using Euclidian distance: Figure 5.9 shows groupings of tags that are close together as measured by Euclidian distance. The same set of symbols □, ○, △, ▽, ◆ and × is in use as in the three-PC score plot in Figure 5.3 and in the hierarchical classification tree using angle measures (Figure 5.8).

The classification tree using Euclidian distance has no major clusters apart from tags 2, 3, 10 and 20, and it does not give the correct results. The reason is that tags with similar spectral features such as those with the □ symbol form plumes rather than spherical clusters, for reasons discussed earlier. In Figure 5.2 tags 2, 3, 10 and 20 are clearly members of the same plume as 24, 25 and 33. However, they are much further from the origin and as a result 24, 25 and 33 lie far away from 2, 3, 10 and 20 in the Euclidian hierarchical tree.

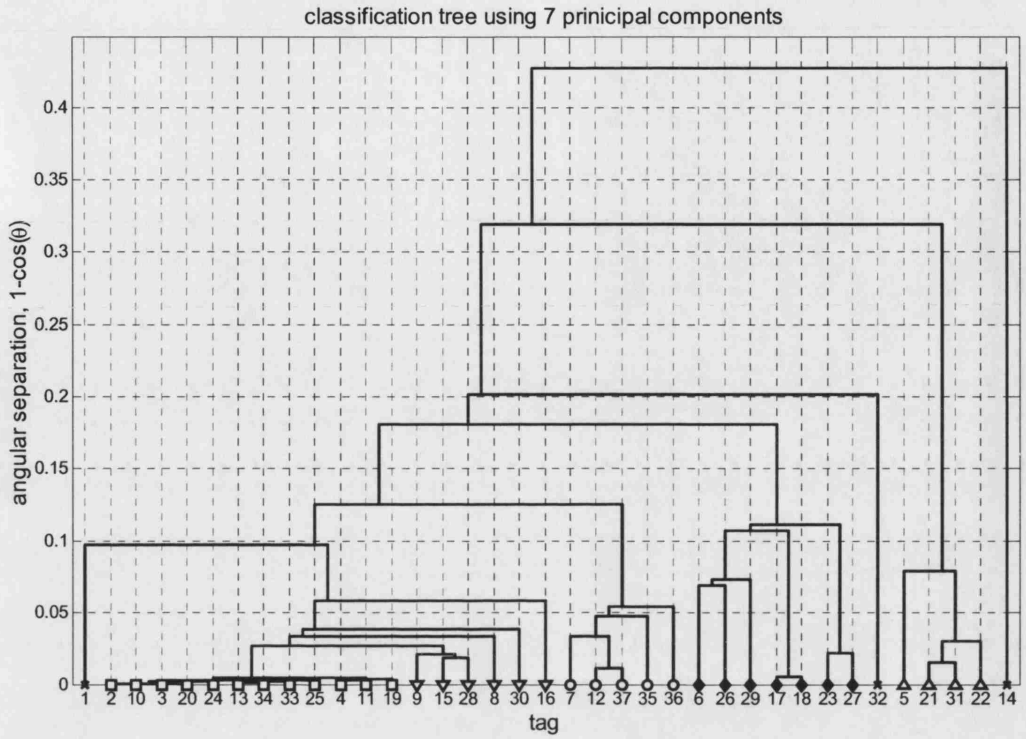

Fig 5.8 Hierarchical classification tree based on angle measures for a seven-PC model. The same symbols have been used as in Figure 5.3.

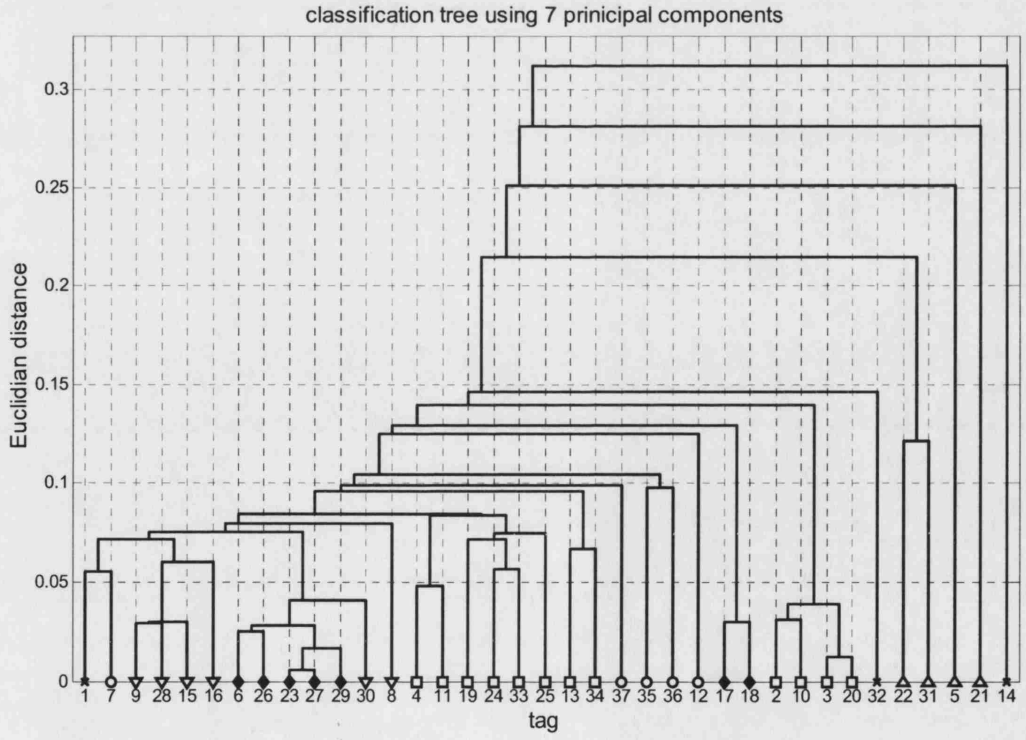

Fig 5.9 Hierarchical classification tree based on Euclidian distance for a seven-PC model.

5.5 REVIEW OF DISTURBANCE DETECTION FINDINGS

5.5.1 Findings from oscillation detection and spectral PCA

Comparisons: Table 5.5 compares the results from basic oscillation detection (i.e. with no filtering) and the clusters of disturbed tags detected by spectral PCA with three and seven principal components. The results from oscillation detection are a sub-set of those from spectral PCA because PCA finds both oscillating and non-oscillating disturbances. Features of note in the comparison are:

- The main upset in the plant, the persistent oscillation at 16.4 samples per cycle is the same in each case with the exception of Tag 28 which was not placed in a cluster by spectral PCA. A close inspection of Figure 5.1 shows the reason; Tag 28 has some other spectral content 4×10^{-2} not shared by the other oscillatory tags;
- Oscillation detection found three oscillatory tags with a oscillation of 85.7 minutes (12, 35, 36). Spectral PCA identified two additional members of the cluster (7 and 37) with similar spectral but whose time trends were not regular enough to be classed as oscillatory.
- Oscillation detection did not identify the cluster 5, 21, 22, 31 with very low frequency deviations;
- Oscillation detection showed that Tags 14 and 32 had their own unique oscillations. Spectral PCA using a seven–PC model was also able to show that Tags 14 and 32 have unique spectra, but a three–PC model could not.

Basic oscillation detection		**Spectral PCA with thee–PC model**		**with seven–PC model**
ave_period	tags	frequency	tags	tags
3.56	32	–	–	32
16.4	2, 3, 4, 8, 9, 10, 11, 13, 19, 20, 24, 25, 28, 33, 34	0.06	2, 3, 4, 10, 11, 13, 19, 20, 24, 25, 33, 34	2, 3, 4, 10, 11, 13, 19, 20, 24, 25, 33, 34
50.9	14	–	–	14
85.7	12, 35, 36	0.012	7, 12, 35, 36, 37	7, 12, 35, 36, 37
–	–	low	5, 21, 22, 31	5, 21, 22, 31

Table 5.5 Comparison of basic oscillation detection, three–PC and seven–PC spectral PCA.

Recommended usage: The findings suggest how oscillation detection and spectral PCA might be used in practice. The aim of a process analyst will be to detect and characterize the main disturbances and to get an estimate of their severity. A sensible way forward would be:

Procedure for detection of distributed disturbances

Step 1: Apply spectral PCA to find and group the main distributed disturbances.

Step 2: Apply basic oscillation detection with no filtering. This step determined whether the groups identified by spectral PCA are oscillatory and if so it finds their period of oscillation;

Step 3: Apply filtering to give the percentage powers. This indicates the extent to which each tag participates in a distributed disturbance or oscillation;

Step 4: Give a recommendation about which oscillations or disturbances should be remedied first.

The aim at this first stage of analysis is to determine which distributed disturbances are having the most severe impact on the plant. Disturbances with high power and with many tags participating will often be the first targets, but process knowledge and insights may suggest starting with the analysis of disturbances affecting the most economically critical loops such as the main throughput flow control or a temperature control that determines the product composition.

The more detailed analysis using multiple oscillation detection and filtering of interfering frequencies is useful in the subsequent investigation stage when it is necessary to discover all the places in the plant having secondary upsets. Section 4.4.2 was able to uncover cases where more than one oscillation affected a tag. For instance, it showed that Tags 4, 11 and 19 were also participating in the 85.7 minute oscillation as well as the 16.4 minute disturbance.

The locations and causes of the disturbances: Figure 5.10 shows the location of some of the disturbed tags on the high level schematic.

A full tag list is available but not all the tag locations were given by the industrial collaborator. Key process measurements are shown but equipment specific tags are not shown, for example Tag 32 whose description is "High Pressure Boiler Feed Water From Pump 7603J" is not shown on this top level schematic.

The 85.4 minute oscillation indicated by O symbols affects the thermodynamic properties of the fuel gas to the feed vaporizer unit. This is clearly a disturbance entering the plant from outside and its origin is in the fuel gas system. The disturbance does not propagate far into the plant, there are no other O symbols in the schematic. It has been damped out either by process dynamics or by disturbance rejection action in a controller.

Only one tag having low frequency deviations could be located in the schematic (△ symbol). Only a few tags are affected by the low frequency deviations and it is not a high priority. Indeed, the industrial collaborator did not know about this disturbance when preparing the data set and schematic.

The schematic shows that the 16.4 minute oscillation indicated by □ symbols is present in many locations and is especially clustered around the reformer and PSA unit. It is this disturbance that the collaborator considered to the highest priority. On first inspection it looks as though Tag 25 is the cause because it is furthest upstream. However, the disturbance is not present in the feed vaporizer or pre-heat sections so the mechanism of propagation from tag 25 is not at all clear. There is a recycle of off-gas from the PSA unit to provide fuel for the reformer, thus a mechanism exists for a disturbance originating in the reformer or PSA unit to become well established, but the results so far do not indicate where in the recycle the disturbance originates. Chapter 6 will present a method for resolving these ambiguities.

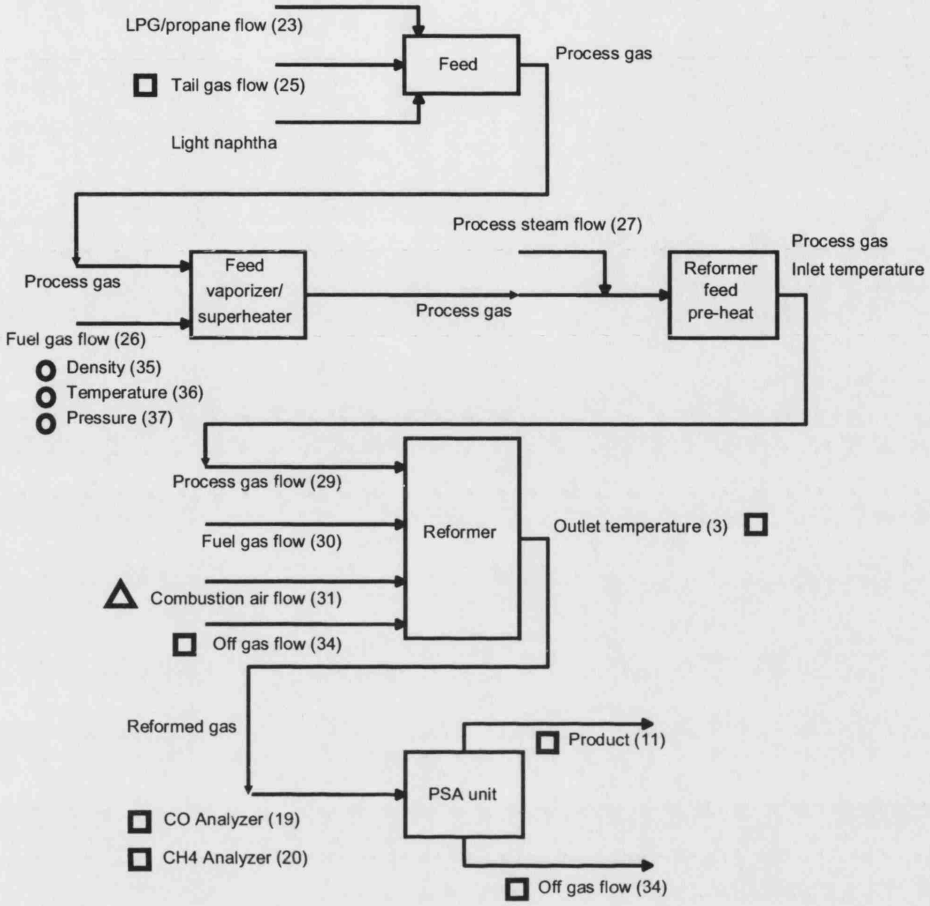

Fig 5.10 SE Asia schematic with the disturbances highlighted. ○ – 85.4 minute oscillation; □ – 16.4 minute oscillation; △ – low frequency deviations.

5.6 CHAPTER SUMMARY

The chapter set out a method called spectral principal component analysis (PCA) for detection of distrusted disturbances that is more general than the oscillation detection method of Chapter 4. Spectral (PCA) detects groups of measurements in a plant that have similar spectral characteristic such as low frequency deviations, broad or overlapping peaks, or sharp spectral peaks that indicate an oscillation.

Spectral PCA gave almost the same results as oscillation detection for oscillatory disturbances but found additional clusters of measurements characterized in one case by low frequency spectral content representing long term deviations from set point and in another by broad-band noise. The tags with broad-band noise were well tuned liquid flow loops operating close to minimum variance.

The chapter discussed the visualization of the results of a PCA analysis. Score plots show how much of each basis vector is present in the spectrum. For instance in the following seven PC model:

$$X = \begin{pmatrix} t_{1,1} \\ \dots \\ t_{m,1} \end{pmatrix} w_1' + \begin{pmatrix} t_{1,2} \\ \dots \\ t_{m,2} \end{pmatrix} w_2' + \dots\dots + \begin{pmatrix} t_{1,7} \\ \dots \\ t_{m,7} \end{pmatrix} w_7' + E$$

the score vector of the i'th spectrum is:

$$s_i = t_{i,1}\hat{z}_1 + t_{i,2}\hat{z}_2 + t_{i,3}\hat{z}_3 + t_{i,4}\hat{z}_4 + t_{i,5}\hat{z}_5 + t_{i,6}\hat{z}_6 + t_{i,7}\hat{z}_7$$

where $\hat{z}_1$ to $\hat{z}_7$ are orthogonal unit vectors in a seven-dimensional space.

If more than three PCs are used then it requires more than three dimensions to visualize the scores. This problem was addressed using a hierarchical classification tree to display the distances between the $s-$ in the high dimensional space. In spectral PCA the clusters of similar tags form plumes radiating from the origin of the score plot rather than spherical clusters. A distance measure based on the angle between the $s-$vectors was successful. Three main plant-wide disturbances were identified by the hierarchical tree in the SE Asia refinery data and three tags were identified that had unique spectral features not shared by any other tag.

Chapter 4 and 5 have demonstrated how oscillating and non-oscillating distributed disturbances can be detected in chemical plants. A recommended procedure has been presented for finding the disturbances and prioritizing them according to their impact on the plant. The benefit of disturbance detection is that it reduces the size of the maintenance task. If a process control engineer finds three distributed disturbances then it is reasonable to assume that there are only three root causes to be tracked down and repaired. The next chapter will present methods to help find the root cause.

6. ROOT CAUSE DIAGNOSIS

6.1 OVERVIEW

6.1.1 Motivation

Disturbances that propagate throughout a plant due to recycle streams, heat integration or other means can have a large impact on product quality and running costs. There is thus a motivation for automated detection of a plant-wide disturbance, as was discussed in Chapters 4 and 5, and for determination of the root cause so that the disturbance may be removed.

Features of interest are:

- An indication of where in the plant the root cause of a plant-wide disturbance is to be found;
- An indication of the nature of the problem;
- A diagnosis that is right first time so that maintenance effort is minimized.

The aim of the root cause diagnosis tools is to pin-point the source of trouble, to indicate the nature of the problem and to provide information-rich statistics to aid experienced engineers in planning of additional tests focused in the right part of the plant.

There are several possible causes of distributed disturbances in chemical plants. This thesis focuses on limit cycles caused by a sticking valve in a feedback loop and thus captures the majority but not all of the causes. Other root causes are currently under investigation in another project and will be commented on at the start of the next section.

6.1.2 Introduction

Chapter 6 of this thesis describes algorithms and software for the purposes of diagnosis of the root cause of a distributed or plant-wide disturbance. A method from the non-linear time-series literature is explained and optimized for use with process measurements when a plant-wide oscillation is present. The focus is on finding the tag or measurement in the plant whose time series has the greatest non-linearity because the measurement with the highest non-linearity is taken to indicate the location of the root cause.

The next section discusses the concepts behind root cause diagnosis and explains why a non-linearity index is an appropriate means to detect the source of a plant-wide disturbance. Section 6.3 optimizes the parameters of the algorithm and gives default values so that it can be applied robustly to new sets of process data without extensive customization. The chapter ends with a worked example which finds the root cause of a distributed disturbance in the SE Asia example of the previous two chapters.

The primary report from the root cause diagnosis tool is a non-linearity index for the time trend of each measurement. At present it is necessary for a user to locate the various measurements in the plant schematic and to manually create a display in the process schematic of where major non-linearity is located. These manually created displays are very useful and deserve further development in future work. One of the recommendations arising from this thesis that there exists a fruitful direction for future research in the automated linkage of data-driven analysis with information about the structure and connectivity of the process.

6.1.3 Work done by the thesis author and others

Work reported in this chapter has been published in the following papers:

> Thornhill, N.F., Cox, J.W., and Paulonis, M., 2003, Diagnosis of plant-wide oscillation through data-driven analysis and process understanding, *Control Engineering Practice*, 11, 1481-1490.
>
> Thornhill, N.F., 2005, Finding the source of a limit cycle oscillation, *IEEE Transactions on Control System Technology, in press*, due for publication March 2005.

The co-authors of the first paper were John Cox and Michael Paulonis of the Eastman Chemical Company, Kingsport, TN, USA. The paper was written by Thornhill who also created the software, results, graphs and all but one of the diagrams. That diagram showing the process schematic was drawn and provided by John Cox. The co-authors also selected a case study which is presented in Chapter 8, provided data, process insights and understanding, and performed additional valve and controller tests in the plant at Kingsport.

6.2 DIAGNOSIS OF ROOT CAUSES

6.2.1 The origins of plant-wide disturbances

Plant-wide disturbances: Following the minimum variance benchmarking approach to controller performance assessment, (Harris, 1989) a *disturbance* is considered to be present when the time series of a plant measurement deviates from white noise. Persistent whole-plant disturbances whether oscillatory and regular or non-periodic are common in control systems in process plants. It is beneficial to track down the root cause of a plant-wide disturbance because then the maintenance effort can be effectively directed towards the equipment or control loop that needs it. The root causes of plant-wide disturbances include:

- Non-linearities such as saturation, dead band, or hysteresis in control valves, sensors or in the process itself that cause limit cycle oscillations;
- Control loop interactions arising when two controllers have a shared mass and/or energy store (e.g. pressure and level controllers may compete for control of the contents of a reactor);
- Structural disturbances caused by coordinated transfers of mass and/or energy between different process units, especially when a recycle is present;
- Disturbances entering at plant boundaries (e.g. due to site utilities such as steam);
- Poorly tuned controllers.

This chapter addresses and solves the diagnosis of non-linear root causes, the first item in the list. The diagnosis of the other root causes is an open and active research area at present. Some suggestions will be given at the end of the chapter for ways forward in future work.

Control loop limit cycles: The motivation for making non-linear root causes a priority is that they are the most common and widespread according to reported surveys (Bialkowski, 1992; Ender, 1993; Desborough and Miller, 2002). A common source of oscillation is a limit cycle caused by a control valve with a deadband or excessive static friction (Åström, 1991; Shinskey, 2000). As shown in chapter 7, control valves with non-linear behaviour such as a dead-band or excessive static friction cause hunting or limit cycling in the control loop. The resulting oscillatory disturbance can upset many downstream units and if a recycle is present the disturbance can also affect measurements upstream of the root cause. A focus upon non-linear root causes can thus be justified because valve friction causes the majority of cases.

Maps of process variable versus manipulated variable (i.e flow through the valve versus valve position) can diagnose the nature of a valve fault and are used in several commercial packages. Other signatures such as odd and even cross correlations and an assessment of the probability density function of the measurement values (Horch, 1999; 2002) have also been used. However, those techniques rely upon a measurement of flow through the valve that is not always present in

an industrial plant. Moreover, without a whole-plant approach to isolate the problem potentially every valve in the plant had to be tested. The aim of this chapter is to enable a precise and rapid diagnosis by means of plant-wide non-linearity testing.

Oscillating disturbances caused by non-linearity in the process itself can also propagate and upset downstream units. A well known example relates to the stop-start nature of flow from a funnel feeding molten steel into a rolling mill (Graebe *et.al.*, 1995). The methods presented in this chapter are able to locate the source of a process non-linearity. For instance, a periodic hydrodynamic instability caused by foaming in a distillation column was described in Thornhill, 2005. Foam in the trays was periodically bubbling up and blocking the column causing a rise in differential pressure which eventually punched through the foam and caused it to fall to the bottom of the column. The column then returned to normal for a spell until the foam built up again.

Propagation of a non-linear limit cycle: Repair of a faulty control loop requires that the engineer knows which control loop should be maintained. In the case of a distributed disturbance it can be a very difficult problem to know which loop to work on because the disturbance from a control loop in a limit cycle typically causes numerous secondary oscillation in other control loops. An automated means is therefore needed to determine which among all the oscillating control loops is the root cause and which are secondary oscillations. Successful studies have used the presence of prominent harmonics to distinguish the source of a limit cycle oscillation from the secondary oscillations in a distillation column in a refinery (Thornhill *et. al.*, 1998) and in a pulp and paper mill (Ruel and Gerry, 1998). The reason why secondary oscillations have lower non-linearity is that as the signal propagates away from its source it passes through physical processes which give linear filtering and which generally add noise. (The filter may be assumed linear if the system is oscillating around a fixed operating point). Such a filter destroys the phase coherence of the time trends and often reduces the magnitudes of the harmonics. Thus non-linearity reduces as the disturbance propagates away from the source and the time trend with the highest non-linearity is thus the best candidate for the root cause.

It is possible that a large amplitude oscillation propagating from a non-linear root cause might drive another part of the process into non-linearity. In that case the non-linearity of the propagated oscillation may rise again for time trends in that part of the plant. An effect of this type was reported with a temperature control loop downstream from the root cause whose valve characteristic had unequal opening and closing characteristics either side of the set point (Thornhill *et. al.*, 2000). The correct interpretation is that the process has two non-linearities, but the root cause diagnosis may then be ambiguous since only one of them is the cause of the disturbance. As discussed in Chapter 8, the process control engineer's insights and knowledge of the process may then be needed to determine the most likely root cause from the candidates presented from data-driven non-linearity analysis.

6.2.2 Introduction to non-linearity diagnosis

Non-linearity in process time trends: A linear time series has a linear dynamic model such as the Box Jenkins model with constant coefficients driven by Gaussian white noise. By contrast, the non-linear time series of interest in this work have a non-linear feedback function:

$$x(n) = \phi\big(x(n-1), x(n-2), ..., u(n-1)\big) + w(n)$$
$$y(n) = h(x(n)) + v(n)$$
$$u(n) = g(y(n))$$

where $\phi\big(x(n-1), x(n-2), ..., u(n-1)\big)$ represents linear dynamics, $x(n)$ is an internal state, $y(n)$ is a measurement from the process, $h(x(n))$ is a measurement function that may be linear or non-linear, $g(x(n))$ is a non-linear feedback function and $w(n)$ and $v(n)$ are process and measurement noises. An example of a non-linear feedback function is the on-off control of a directly-injected steam heated tank in which the steam valve switches on when the temperature drops to a low limit and switches off again when the temperature reaches a high limit. The non-linear characteristic g in that case is a relay with deadband. The temperature is not steady in such a system and it cycles in periodic pattern. Control loops having valves with non-linear friction characteristics and non-linear instrumentation faults can lead to similar behaviour and are described by the same model structure.

The waveform of a limit cycle is periodic but non-sinusoidal and therefore has harmonics. A distinctive characteristic of a nonlinear time series is the presence of phase coupling which creates coherence between frequency bands such that the phases are non-random and form a regular pattern. Nonlinearity may thus be inferred from the presence of harmonics and phase coupling.

Nonlinearity detection: Methods for nonlinearity detection in time series include those using surrogate data (Theiler *et. al.*, 1992; Kantz and Schreiber, 1997) which have been used in applications ranging from analysis of EEG recordings of epileptic people (Casdagli *et. al.*, 1996) to the analysis of x-rays emitted from a suspected astrophysical black hole (Timmer *et. al.*, 2000). Surrogate data are times series having the same power spectrum as the time series under test but with the phase coupling removed by randomization of phases. A key property of the test time series is compared to that of its surrogates and nonlinearity is diagnosed if the property is significantly different in the test time series.

Another method of non-linearity detection uses higher order spectra because these are sensitive to certain types of phase coupling. For instance, the bispectrum (Rao and Gabr, 1980; Hinich, 1982; Nikias and Petropulu, 1993) responds to quadratic phase coupling as discussed in Chapter 2.

Bispectrum and bicoherence have been used to detect the presence of nonlinearity in process data (Emara-Shabaik *et. al.*, 1996; Choudhury *et. al.*, 2004). A disadvantage of the bispectrum for detection of nonlinear limit cycles is that the oscillations may have symmetrical waveforms (e.g.

a square wave or triangular wave) and the bispectrum of a symmetrical waveform is zero. Zang and Howell (2003) have investigated the types of limit cycles that are amendable to bispectrum analysis.

The presence of harmonics in a time series has also been used successfully for diagnosis of SISO control loop faults (Paulonis and Cox, 2003; Thornhill and Hägglund, 1997; Harris *et. al.*, 1996). Finding harmonics requires signal processing to isolate the spectral frequencies of interest and inspection to confirm that the frequencies are integer multiples of a fundamental. The inspection is often undertaken by visual examination of the spectra and is therefore unsuitable for a large-scale implementation involving several hundred or even a thousand or more plant measurements. Moreover, it is possible that components at the harmonic frequencies are not phase coupled in which case the harmonic signature will be a misleading indicator of nonlinearity.

Other quantities from non-linear time series analysis such as correlation dimension and maximal Lyapunov exponent are also being examined in the literature (Zang and Howell, 2004). These are promising because their numerical values are often different for the primary disturbance compared to the secondary disturbances. They have not yet had detailed investigation, exhaustive testing nor have their parameters been optimized, however. Bicoherence is also being examined for plant-wide root cause diagnosis at the University of Alberta although the work has not been published at the time of writing.

6.2.3 Diagnosis of non-linearity

Non-linearity detection using surrogates: The non-linearity test described in this sub-section determines whether a time series could plausibly be the output of a linear system driven by Gaussian white noise, or whether its properties can only be explained as the output of a non-linear system (Theiler *et. al.*, 1992; Kantz and Schreiber, 1997; Schreiber and Schmitz, 2000). The underlying concept is that non-linearity reduces as the disturbance propagates away from the root cause. Therefore the root cause is to be found in the part of the plant where the non-linearity of the time series of the plant measurements is highest.

A time series with phase coupling is more structured and more predictable than a similar time series known as a *surrogate* having the same power spectrum but with random phases (Theiler *et. al.*, 1992). The spread of values of some statistical property of a group of surrogate data trends provides a reference distribution against which the properties of the test time series can be evaluated.

Predictability: The basis of the test is a comparison of the predictability of the time trend compared to that of its surrogates. Figure 6.1 illustrates the concept. The top panel is an oscillatory time trend of a steam flow measurement from a refinery. It has a clearly defined pattern and a good prediction of where the trend will go after reaching a given position, for

example at one of ringed peaks, can be achieved by finding similar peaks in the time trend and observing where the trend went next on those occasions.

The middle panel shows a surrogate of the time trend. By contrast to the original time trend the surrogate lacks structure even though it has the same power spectrum. The removal of phase coherence has upset the regular pattern of peaks. For instance, it is hard to anticipate where the trajectory will go next after emerging from the region highlighted with a circle because there are no other similar peaks on which to base the prediction.

The lower two panels introduce the concept known as *embedding*, a trajectory of the time trend plotted against a time-delayed version of the same signal. The lower left plot confirms the predictable and structured nature of the steam flow time trend. There are large regions of the embedded space that are never visited by the trajectory. The white circle corresponds to the peak positions and it is easy to predict where the trajectory will go next after visiting the circle because it always does the large vertical downwards jump on the right hand side. The embedded plot in the lower right panel is for the surrogate and it is much less structured. It is not possible to predict how the trajectory will enter or leave the white circle because the entry and exit points on previous occasions were random. The time delay of three samples in these embedded plots was selected for the purposes of demonstration. Both the dimensions and the delays used in the embedded space are key parameters in the algorithm and their optimization will be discussed in section 6.3.

Predictability of the time trend relative to the surrogate gives the basis of a non-linearity measure. Prediction errors for the surrogates define a reference probability distribution under the null hypothesis. A non-linear time series is more predictable than its surrogates; a prediction error for the test time series which is smaller than the mean of the reference distribution by more than three standard deviations suggests the time trend is non-linear.

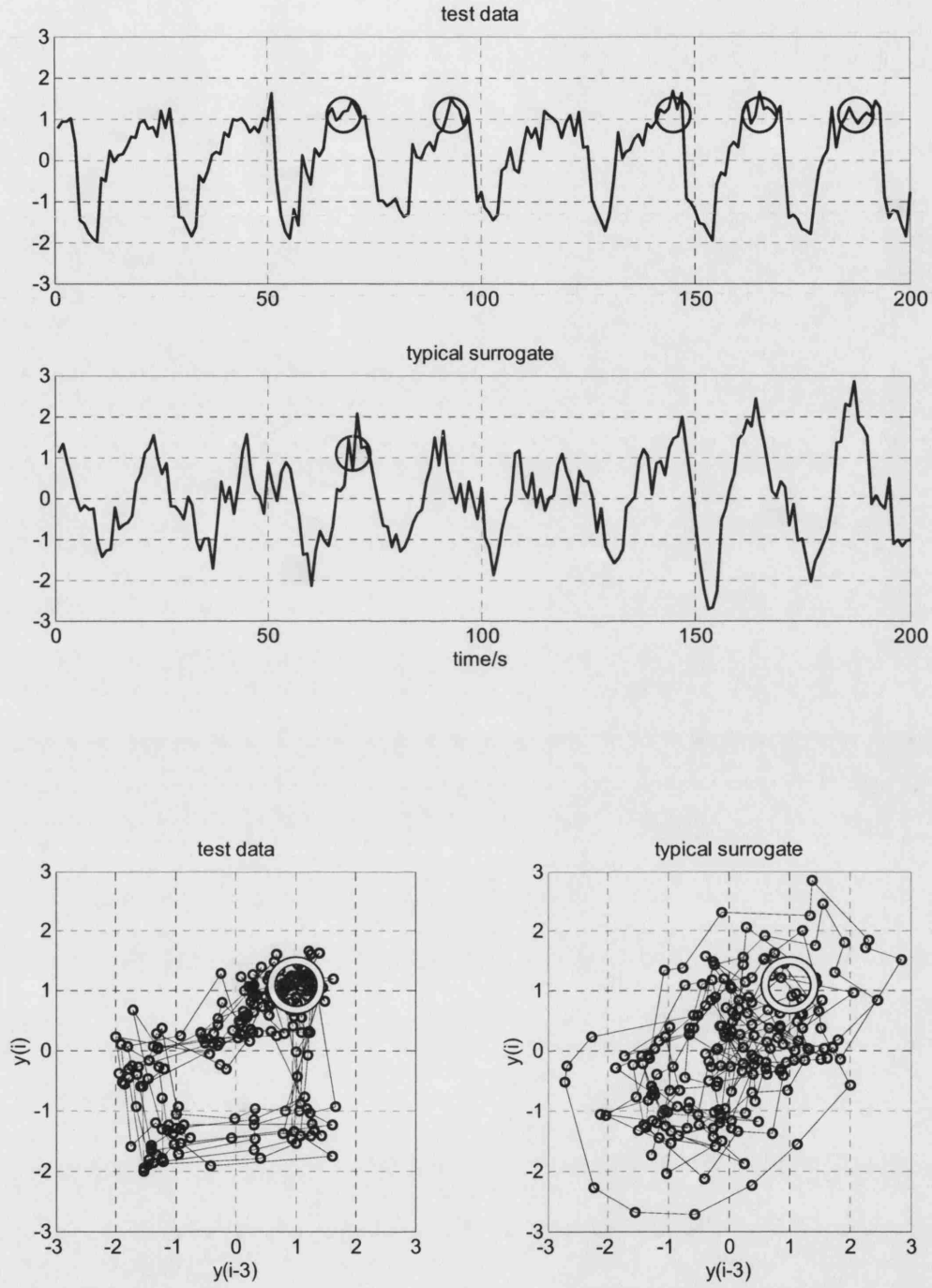

Fig 6.1 Upper panels: A limit cycle oscillation and a typical surrogate represented as time trends. Lower panels: Limit cycle and surrogate as a two-dimensional embedding. The white rings in the embedded plots indicate the positions of the peaks circled in the time trends.

6.2.4 Algorithm for non-linearity diagnosis

Construction of the data matrix: Non-linear prediction of time series was described by Sugihara and May (1990) to distinguish determinism from random noise, and the field of non-linear time series analysis and prediction has been reviewed by Schreiber (1999). Rhodes and Morari (1998) gave an early process application of non-linear prediction where the emphasis was on modelling of non-linear systems when noise corrupts a deterministic signal.

Non-linear prediction uses a data matrix called an embedding having E columns each of which is a copy of the original data set delayed by a suitable sampling interval. For instance, a data matrix with a delay of one sampling interval and $E = 3$ is:

$$\boldsymbol{Y} = \begin{pmatrix} y(1) & y(2) & y(3) \\ y(2) & y(3) & y(4) \\ y(3) & y(4) & y(5) \\ \cdots & \cdots & \cdots \\ y(\ell-2) & y(\ell-1) & y(\ell) \end{pmatrix}$$

Rows of the matrix $\boldsymbol{Y}$ represent time trajectories that are segments of the original trend. Since the original data formed a continuous time trend the trajectories in adjacent rows are similar. They are called *near-in-time neighbours*. Also, if the time trend is oscillatory then the trajectories in later rows of $\boldsymbol{Y}$ will be similar to the earlier rows after one or more complete cycles of oscillation. For instance, if the period of oscillation is 50 samples per cycle then $\|\underline{y}_{51} - \underline{y}_1\|$ will be small, where $\underline{y}_{51}$ is the 51'st row vector of $\boldsymbol{Y}$ and $\underline{y}_1$ is the first. Those rows are called *near neighbours*.

Calculation of prediction error: Predictions are generated from near neighbours. Near-in-time neighbours are excluded so that the neighbours are only selected from other cycles in the oscillation. When k nearest neighbours have been identified then those near neighbours are used to make an H step-ahead prediction. For instance, if row vector $\underline{y}_1$ were identified as a near neighbour of $\underline{y}_{51}$ and if H were 3 then $y(4)$ would give a prediction of $y(54)$. A sequence of prediction errors can thus be created by subtracting the average of the predictions of the k nearest neighbours from the observed value. The overall prediction error is the r.m.s. value of the prediction error sequence.

The analysis is non-causal and any element in the time series may be predicted from both earlier and later values. Figure 6.2 illustrates the principle using a time series from the SE Asia refinery case study where the embedding dimension E is 16 and the prediction is made 16 steps ahead. The upper panel shows the 100'th row of the data matrix $\boldsymbol{Y}$ which is a full cycle starting at sample 100, marked with a heavy line. Rows of $\boldsymbol{Y}$ that are nearest neighbours of that cycle begin at samples 67, 133, 166, 199, and 232 and are also shown as a heavy lines in the lower panel. The average of the points marked O in the lower panel are then used as a prediction for the value marked ×.

Data pre-processing: The periodic nature of the detected oscillation may be exploited in order to give robust default settings for the parameters. A summary list is presented here and the detailed reasoning behind the recommendations will be presented in Section 6.3. With the data pre-processing steps indicated here the default parameters can be used for any oscillating time trend.

Data pre-processing

Step 1: The period of the plant-wide oscillation is determined ;

Step 2: The number of samples per cycle, S, is adjusted to be no more than 25. The time trends are sub-sampled if necessary;

Step 3: The number of cycles of oscillation in the data set should be at least 12 for a reliable non-linearity estimate;

Step 4: The selected data are end-matched to find a subset of the data containing an integer number of full cycles. The algorithm and other issues associated with end-matching are explained in detail in section 6.3.2.

Step 5: The end-matched data are mean centred and scaled to unit standard deviation. The sequence $y(1) \ldots . y(\ell)$ denotes end-matched and pre-processed data in the following sections.

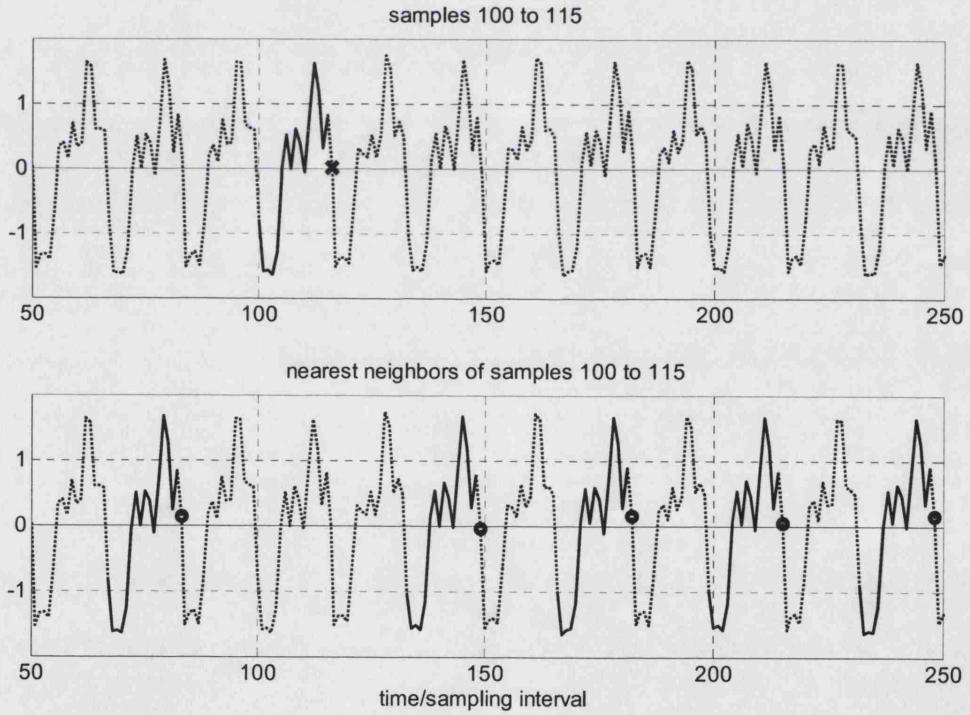

Fig 6.2 Illustration of the nearest neighbour concept. The highlighted cycles in the lower plot are the five nearest neighbours of the cycle in the upper plot. The average of the points marked ○ gives a prediction for the point marked ×.

Derivation of surrogate data: Surrogate data are derived from the pre-processed and end-matched time trend. Surrogate data have the same power spectrum as the time trend under test. The magnitudes of the discrete Fourier transform (DFT) are the same in both cases but the

arguments of the DFT of the surrogate data set are randomized. Thus if the DFT in frequency channel i is:

$$|Y(j\omega_i)|e^{j\angle Y(j\omega_i)}$$

then the DFT of the surrogate is:

$$|Y(j\omega_i)|e^{j(\angle Y(j\omega_i)+\phi_i)}$$

where ϕ_i is a phase selected from a uniform random distribution in the range $-\pi < \phi_i \leq +\pi$. The aliased frequency channels above the Nyquist sampling frequency have the opposite phase added. If the number of samples ℓ is even and if the frequency channels are indexed as $i=1$ to ℓ the Nyquist frequency is in channel $\ell/2+1$ and the alias of the i'th frequency channel is channel $\ell-i+2$. Then:

$$\phi_1 = 0,\ \phi_{\ell/2+1} = 0 \text{ and } \phi_{\ell-i+2} = -\phi_i \ (i = 2 \text{ to } \ell/2)$$

If ℓ is odd:

$$\phi_1 = 0 \text{ and } \phi_{\ell-i+2} = -\phi_i \ (i = 2 \text{ to } ceil(\ell/2))$$

where $ceil(\ell/2)$ is the rounded-up integer value of $\ell/2$.

Finally, the surrogate data set is created from the inverse Fourier transform of the phase randomized DFT.

Statistical testing: The non-linearity test requires the determination of mean square prediction errors of M surrogates. The statistical distribution of those errors gives a reference distribution. If the test data prediction error lies on the lower tail of the reference distribution then the test signal is more predictable and non-linearity is diagnosed using the following statistic based on a three-sigma test:

$$N = \frac{\bar{\Gamma}_{surr} - \Gamma_{test}}{3\sigma_{\Gamma_{surr}}}$$

where Γ_{test} is the mean square error of the test data, $\bar{\Gamma}_{surr}$ is the mean of the reference distribution and $\sigma_{\Gamma_{surr}}$ its standard deviation. If $N>1$ then non-linearity is inferred in the time series. Larger values of N are interpreted as meaning the time series has more non-linearity, those with $N<1$ are taken to be linear.

It is possible for the test to give small negative values of N. Negative values in the range $-1 \leq N < 0$ are not statistically significant and arise from the stochastic nature of the test. Results giving $N<-1$ do not arise at all because the surrogate sequences which have no phase coherence are always less predictable than a non-linear time series with phase coherence.

Algorithm summary:

Algorithm: Non-linearity detection using surrogate data

Step 1. Form the embedded matrix from a preprocessed and end matched subset of the test data $y(1) \ldots y(\ell)$:

$$\mathbf{Y} = \begin{pmatrix} y(1) & y(2) & \ldots & y(E) \\ y(2) & y(3) & \ldots & y(E+1) \\ y(3) & y(4) & \ldots & y(E+2) \\ \ldots & \ldots & \ldots & \ldots \\ y(\ell-E+1) & y(\ell-E+2) & \ldots & y(\ell) \end{pmatrix}$$

Step 2. For each row $\underline{\mathbf{y}}_i$ of $\mathbf{Y}$ find the indexes j_p $(p = 1 \ldots k)$ of k nearest neighbour rows $\underline{\mathbf{y}}_{j_p}$ having the k smallest values of $\left\| \underline{\mathbf{y}}_{j_p} - \underline{\mathbf{y}}_i \right\|$ subject to a near-in-time neighbour exclusion constraint $\left| j_p - i \right| > E/2$.

Step 3: Find the sum of squared prediction errors for the test data:

$$\Gamma_{test} = \sum_{i=1}^{\ell-H} \left(y(i+H) - \frac{1}{k} \sum_{p=1}^{k} y(j_p + H) \right)^2$$

Step 4: Create M surrogate prediction errors Γ_{surr} by applying steps 1 through 3 to M surrogate data sets.

Step 4: Calculate the non-linearity from

$$N = \frac{\bar{\Gamma}_{surr} - \Gamma_{test}}{3\sigma_{\Gamma_{surr}}}$$

6.3 OPTIMIZATION OF PARAMETERS

6.3.1 Default parameter values

Recommended default parameter values: Empirical studies have been carried out to ascertain the sensitivity of the non-linearity index to the parameters of the algorithm. Reliable results have been achieved using the default values shown in Table 6.1. The *floor* function in the third row indicates that for non-integer values of S then E is set to the rounded-down integer value of S.

description	value
samples per cycle, S	$7 \le S \le 25$
number of columns in embedded matrix, E	$E = floor(S)$
prediction horizon, H	$H = E$
number of cycles of oscillation, C	$C \ge 12$
number of near neighbours, k	$k = 8$
number of surrogates, M	$M = 50$

Table 6.1 Suggested default values for the parameters of the non-linearity detection algorithm.

The next sub-sections explore each one of these recommendations showing why they were chosen. Selected oscillatory time trends from the SE Asia data set were used for the evaluation. Figure 6.3 shows mean centred data normalized to unit standard deviation and the spectra are scaled to the same maximum peak heights.

The time series of the measurements in the top three rows of Figure 6.3 are non-linear and are close to the root cause. Their spectra have harmonics and the phase patterns are not random. Those in the bottom two rows are far from the root cause and are linear. The data have an oscillation period of 16.7 sampling intervals and the conditions used were varied around default values of $C = 12$, $E = H = 16$, $k = 8$ and $M = 50$.

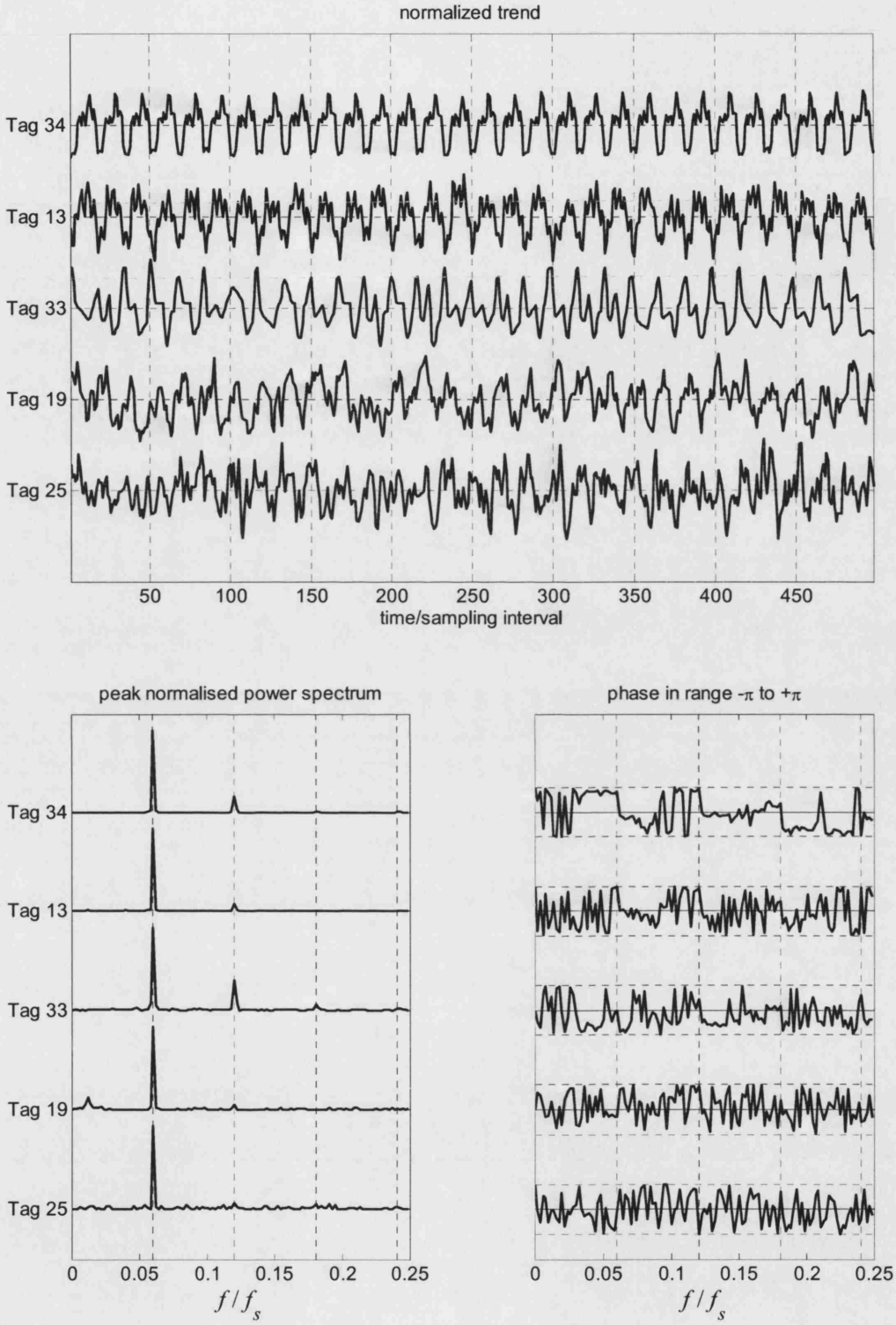

Fig 6.3 Time trends and spectra of the data used for detailed evaluation.

Default number of samples per cycle: It is practical to limit the number of samples per cycle, S. There is a trade-off between the number of samples needed to properly define the shape of a non-sinusoidal oscillation on the one hand and the speed of the computation on the other. The algorithm requires a distance measure to be ascertained between every pair of rows in the embedded matrix and therefore the time taken for the computation increases as Q^2, where $Q = C \times S$ is the total number of samples in the time trend. Therefore the number of samples per cycle, S, and the number of cycles, C, cannot be increased arbitrarily. Data sets of 418 samples ($S = 16.7$ and $C = 25$) gave successful results in the SE Asia case study reported in section 6.3.3. The computation time with a 1.2GHz Pentium processor using Matlab 6.5 was about 30s per tag. In general it is recommended to keep the size of the data set below about 500 in an industrial implementation so that results are produced in a reasonable time.

It would be infeasible to operate with fewer than 7 samples per cycle because harmonics would not be satisfactorily captured. With $S = 7$, any third harmonic present is sampled at 2.33 samples per cycle which just meets the Nyquist criterion of 2 samples per cycle. The reason for focusing the recommendation on the third harmonic is that it is the most prominent harmonic in symmetrical non-sinusoidal oscillations having square or triangular waveforms.

There is an extensive literature on the topic of how to choose the sampling interval correctly. The considerations apply to the columns of the $\boldsymbol{Y}$ matrix in section 6.2.4. What is needed is for adjacent columns to be "somewhat independent but not completely uncorrelated" (Lai and Lerner, 1998). If the sampling interval is too small the $\boldsymbol{Y}$ matrix becomes packed with redundant columns that do not contribute to the analysis. The upper limit of 25 samples per cycle was selected on the basis that 25 samples is enough to accurately capture the waveforms of the oscillations observed in practical case studies. By contrast, if 50 or 100 samples per cycle were used then adjacent columns would be almost identical apart from random noise, and the computation time would increase greatly because it would still be necessary to include an adequate number of complete cycles of oscillation.

Default embedding dimension and prediction horizon: Figure 6.4(a) shows the effect of changing E and H. They were kept equal to each other and both were varied together. The threshold of non-linearity ($N = 1$) is also shown in the plot (horizontal dashed line) as well as the $E = H = 16$ default for the data set (vertical dashed line). Once E becomes larger than half a cycle of the oscillation, in this case when $E > 8$, the results for the non-linearity index N become quite steady while for small values of E the index falls towards the $N = 1$ threshold.

An aim of the work presented here is to give reliable default values that are easy to determine. Determination of the period of oscillation S is becoming a standard component of controller performance tools (Matrikon Inc., 2004; Expertune Inc., 2004). Therefore the recommendation to set $E = floor(S)$ is robust because it is in the steady region of Figure 6.4(a) and is easy to implement because S is already known.

The poor performance with small values of E arises because of the phenomenon of false near neighbours (Rhodes and Morari, 1997), especially when the time trends have high frequency

features or noise. The upper panel in Figure 6.5 shows an example of what can happen when the embedding dimension is small, in this case $E = 2$. The row of the **Y** matrix starting at sample 159 comprises just two samples, 159 and 160, shown as small square symbols. Near neighbours are shown in the lower panel, these are two-sample segments of the time trend whose values are similar to samples 159 and 160. However, some of them are false neighbours because they are not from matching parts of the trend. The average of the points marked ○ are used as a prediction for the value marked ×, but some of them such sample 62 which is based on a false neighbour are not accurate.

Default number of cycles and near neighbours: Figure 6.4(b) (top right panel) shows a plot of the number of cycles of oscillation presented for analysis versus the non-linearity index. The value of the non-linearity statistic fluctuates somewhat but when the number of cycles becomes too few the results start to become very unreliable. For instance, when more than twelve cycles are used in the analysis then tag 33 is consistently and correctly reported as non-linear but when fewer than 12 cycles are present its non-linearity index and that of other tags drops towards the $N = 1$ threshold. On the basis of these examples it seems necessary to use 12 cycles of oscillation or more. The non-linearity index becomes consistent if this condition applies and the ranking order of the tags is maintained, for instance tag 34 consistently has the highest non-linearity index if $C \geq 12$.

Figure 6.4(c) (lower left panel) fixes the number of cycles of oscillation at 12 and varies k, the number of near neighbours used for prediction. The results are quite steady over a wide range of values of k although some of the tags with non-linearity show a drop towards the $N = 1$ threshold for large k. A recommended value for k can be based on the number of cycles of oscillation. From common sense reasoning, it is sensible to make sure that number of near neighbours is smaller than the number of cycles because each near neighbour is one whole cycle if the $E = S$ recommendation is adopted. Given that one or two cycles may be lost during end-matching a conservative choice is $k = 8$ when the number of cycles is 12. The same conservative reasoning suggests that for other cases k should never be greater than $C - 4$ although in practice it has been found quite satisfactory to just set the value to $k = 8$.

The reason why the non-linearity tests gives less reliable results for large k when the number of cycles, C, is fixed at 12 is that the useful near neighbours run out. For instance, if $k = 20$ and if the data set has twelve cycles then any one cycle has eleven near neighbours that are closely matching cycles starting at the same position in the oscillation, like those in Figure 6.2. The remaining near neighbours will have to be selected from other parts of the oscillation and will not be such good matches.

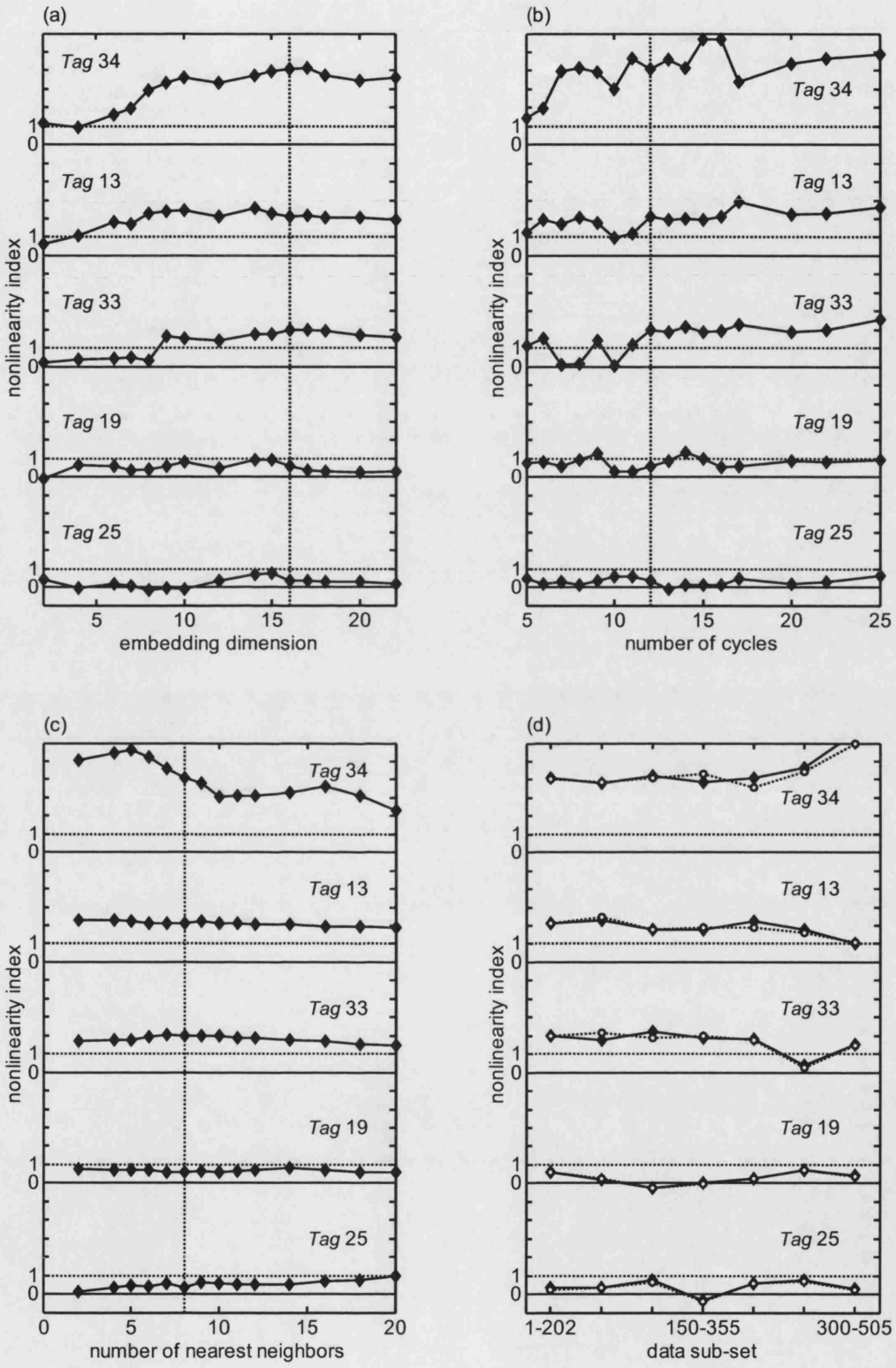

Fig 6.4 Effects of parameters of the algorithm on the calculated non-linearity index. The vertical dashed lines in panels (a) to (c) show the recommended default values.

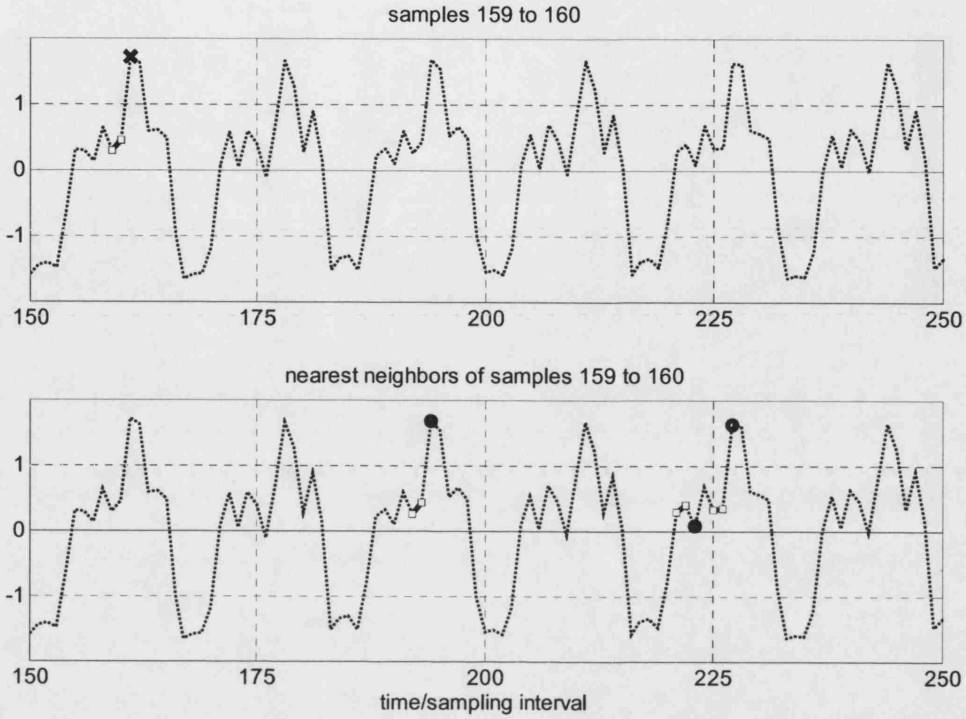

Fig 6.5 Illustration of false near neighbours when $E = 2$. The samples at 221 and 222 are similar to those at 159 and 160 but are from a different part of the cycle. A prediction of sample 161 based on sample 223 will be inaccurate.

Assessment of variability of the index: Figure 6.4(d) (lower left panel) shows the variability in the non-linearity index as the data sub-set is altered. The data set was divided into seven overlapping subsets each having 12 cycles of oscillation. The first subset comprised samples 1 to 202, the second was 50 to 252, and so on. Figure 6.4(d) shows that different parts of the data trend exhibit varying amounts of non-linearity. It is noted, however, that the same conclusion about the tag with the highest non-linearity applies regardless of the data subset. Tag 1 (34) has the highest non-linearity right across the board, tags 13 and 33 are next highest and 19 and 25 have little non-linearity.

The white dots in Figure 6.4(d) show the effect of varying the number of surrogates. Since it is a statistical test there must be enough surrogates to properly define the reference distribution. The difference between the results from 50 surrogates (black diamonds) and 250 surrogates (white circles) is about ± 0.1 overall and is less than the variability caused by the data sub-set. It is therefore concluded that 50 surrogates are enough.

6.3.2 End matching

False-positive results with strongly cyclic data: Issues have been identified with the use of surrogate data with cyclic time series (Stam *et. al.*, 1998; Small and Tse, 2002). The surrogate is derived by taking the discrete Fourier transform (DFT) of the test data, randomization of the arguments followed by an inverse DFT. Nonlinearity testing based on strongly cyclic data can give rise to false detection of non-linearity because when the time trend is strongly cyclic then artifacts in the DFT due to end-matching effects influence the surrogates. Unless care is taken with the end matching the test may give false positive results and report a non-zero non-linearity index for a linear time series.

The reason for the false positive results is the phenomenon of spectral leakage in the discrete Fourier transform caused by the use of a finite length data sequence. Figure 6.6 illustrates the effect of spectral leakage. The upper panel shows the DFT of a sine wave having 8 samples per cycle when the total data length is an exact multiple of the period, in this case exactly 8 cycles. The DFT is zero in all frequency channels expect the one at $f/f_S = 0.125$ corresponding to the frequency of the oscillation. By contrast, the lower panel shows the DFT when the total data length is a complete number of cycles minus one sample. It has a non-zero magnitude in frequency channels adjacent to the channel containing the main spectral peak. A phase randomized surrogate derived from the DFT in the lower panel therefore contains frequencies that were not present in the original signal and will thus be less predictable than the original sine wave giving a false indication of non-linearity. The true surrogate of a sine wave is a phase shifted sine wave at the same frequency and is equally predictable.

It is therefore necessary to take special precautions when analyzing cyclic time series. Stam *et. al.* (1998) used an end-matching step that ensures the data length of the time series is an exact multiple of the period of the cycle to avoid false non-linearity detection. Small and Tse (2002) proposed the calculation of special constrained surrogates that pay particular attention to frequencies in the data set having periods longer than the period of the strong cycles. That solution is not applicable to industrial data where the non-linearity of interest is distortion of the periodic waveform, however.

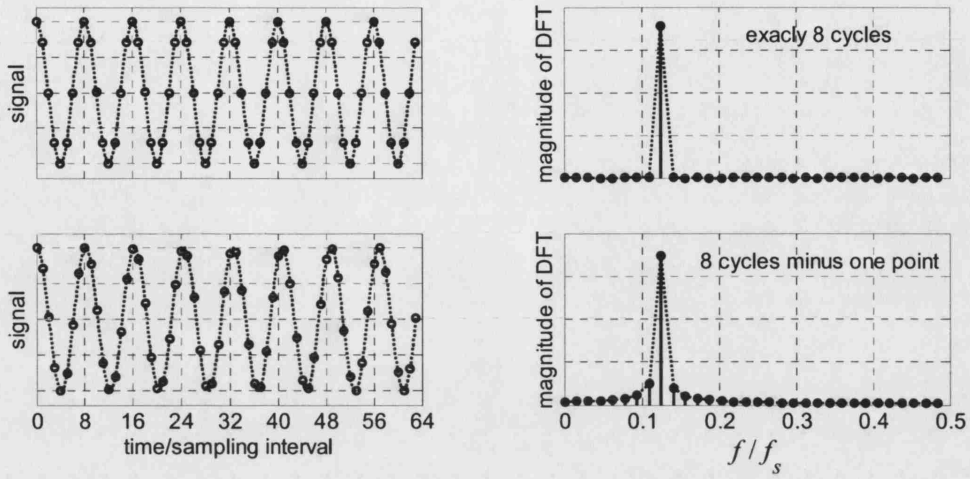

Fig 6.6 Illustration of the importance of end matching. For a strongly cyclic time trend the data set should be an exact number of cycles of oscillation otherwise the Fourier transform will give spectral leakage into adjacent frequency channels.

The end-matching criterion: Surrogate data analysis requires a sub-set of the data such that the starting gradient and value match well to the final gradient and value. Hegger *et. al.*, (2000) recommend finding a sub-set of the non-matched data (denoted by samples x) with n samples starting at x_i and ending at x_{i+n-1} which minimizes the sum of the normalized discontinuities $\left(d_0 + d_1\right)$ between the initial and end values and the initial and end gradients, where:

$$d_0 = \frac{\left(x_i - x_{i+n-1}\right)^2}{\sum_{j=i}^{i+n-1}\left(x_j - \bar{x}\right)^2}$$

and

$$d_1 = \frac{\left(\left(x_{i+1} - x_i\right) - \left(x_{i+n-1} - x_{i+n-2}\right)\right)^2}{\sum_{j=i}^{i+n-1}\left(x_j - \bar{x}\right)^2}$$

with $\bar{x}$ being the mean of the sequence x_i ..... x_{i+n-1}.

End matching of an oscillating time trend as described above creates a time trend where the last value is the first sample of another cycle. An end matched sequence which contains an exact number of cycles is x_i ..... x_{i+n-2} and is derived from the x_i ..... x_{i+n-1} sequence by omitting the last sample.

Application to industrial data: Experimental laboratory data will give problems of the type outlined above when the experimental system is driven by a cyclic source such as a laboratory signal generator. This is termed in the literature as *strongly cyclic*. Industrial data, however, even

when cyclic do not often suffer from the problems described above because the cyclic behaviour is not normally "strong". Although they might be readily detectable the cycles are normally not completely regular and the spectral power in the test signal is already is spread across several frequency channels. A good example is Figure 6.7 showing oscillatory data from a separation column in an Australian refinery from Thornhill *et al.*, (1996*b*). The time trend labelled AC1 has two shorter cycles at around samples 230 to 250 and two longer cycles between samples 100 and 150. The effects of spectral leakage in the DFT are less severe in this signal because a wider range of frequency channels are already occupied. Therefore the surrogates more accurately represent the real signals than in the case of strong cyclic data.

Table 6.2 shows non-linearity calculations for AC1 and a pure sine wave of the same period of oscillation as the average of AC1. The correct result for the sine wave is $N<1$, a result which is achieved when the subset of the data is an exact number of cycles of oscillation. If the subset is longer or shorter, even by one sample, there is a false non-linearity detection because of spectral leakage contaminating the surrogates. By contrast, industrial data such as AC1 with its less regular cycles are not sensitive to minor variations in the end-match. No false non-linearity was detected for AC1 even when mismatched at the ends by one or two samples.

end matching	*N* for sine wave	*N* for industrial data AC1
one point extra	**4.72**	0.89
end matched	0.11	0.86
one point fewer	**5.07**	0.83
two points fewer	**5.19**	0.84

Table 6.2 The effect of end matching on false-positive results with synthesized and industrial data. The synthesized sine wave gives false-positives ($N>1$) when the end matching is not exact.

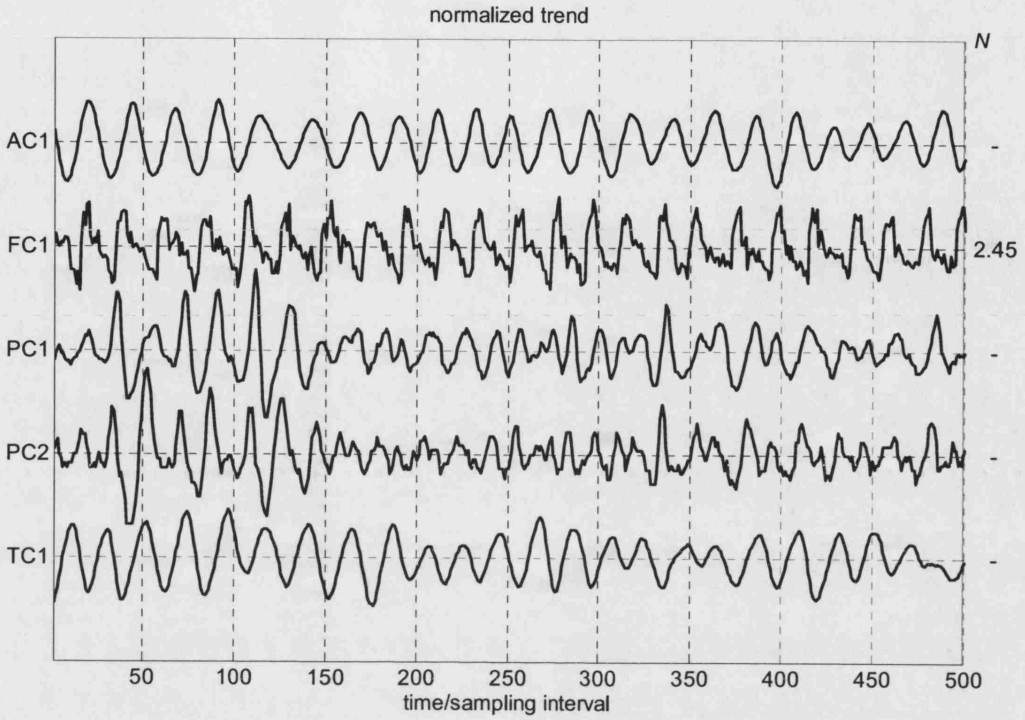

Fig 6.7 Time trends from an industrial study in Thornhill *et. al.*, 1996*b*. Non-linearity indexes greater than one are shown on the right. The trends are cyclic but not *strongly cyclic*.

6.4 WORKED EXAMPLE AND DISCUSSION

6.4.1 Application to the SE Asia data set

Plant-wide oscillation caused by a valve fault: Non-linearity diagnosis has been applied to the data set of the hydrogen reformer from a SE Asian refinery. The measurements from this plant have been discussed in Chapters 4 and 5 where the main disturbance was shown by spectral principal component analysis to be a 16.7 minute oscillation in the reformer and PSA unit.

The process control engineer for the plant has emphasized that the pressure cycle swings of the PSA unit were not the cause. The root cause was known to be a valve fault in the recycle of fuel gas from the PSA unit to the reformer. The aim of the analysis is to determine whether the non-linearity test delivers the correct diagnosis by focusing attention on measurements closest to the root cause. The average oscillation period was 16.7 samples, the settings for the algorithm were therefore $E = H = 16$ and $k = 8$. A data set comprising 25 cycles of oscillation was used.

Figure 6.8 shows mean centred normalized controller errors (1 minute samples) with the non-linearity indexes given on the right hand side. The largest non-linearity index is for tag 34. Therefore the flow measured by tag 34 is identified as the measurement closest to the physical root cause.

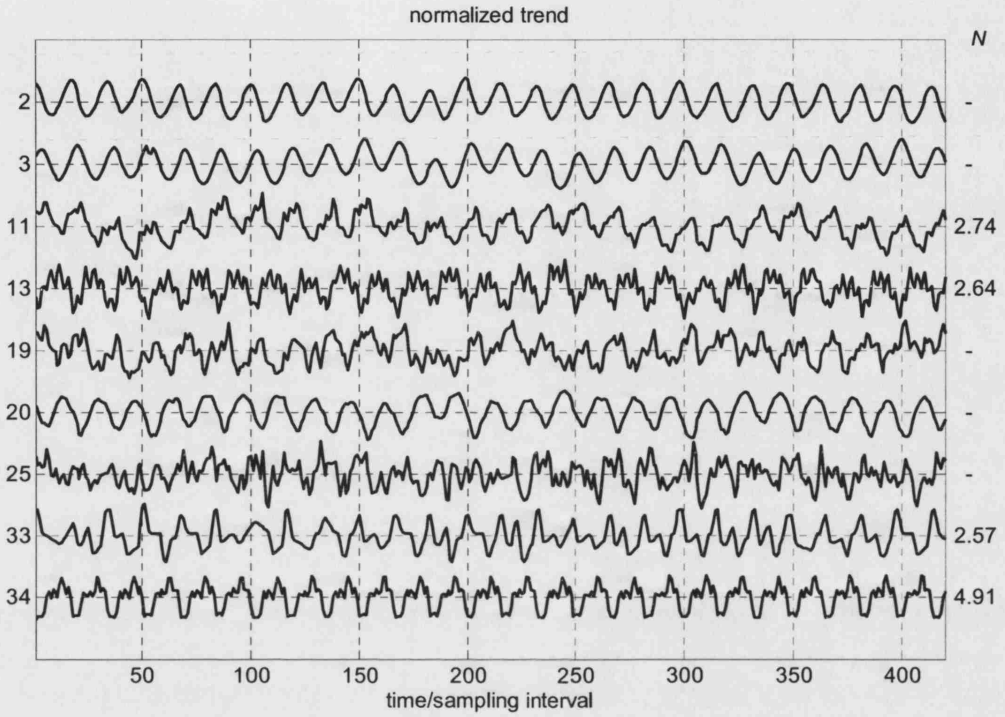

Fig 6.8 Time trends of oscillating tags from SE Asia data set. The trends are for the controller errors. Non-linearity indexes greater than one are shown on the right.

The diagnosis of this process is a challenge because Tag 34 is in the PSA unit which is the last unit in the plant from where the final product leaves the plant, as shown in Figure 6.9. The mechanism for propagation to the reformer unit is the off-gas from the PSA unit which is used as fuel gas for the reformer. The oscillatory upset in the PSA unit propagates to the reformer because the flow of fuel gas varies.

Tag 25 is upstream yet it was still influenced by the oscillation, as can be seen in the autocovariance plots of Figure 4.6. Inspection of the flow sheet alone might misleadingly suggest that Tag 25 is the root cause because it is furthest upstream. Such a theory is hard to defend, however. It would require an explanation of how Tag 25 can cause oscillations in the reformer and PSA unit without upsetting the feed vaporizer and pre-heat sections where no oscillations were detected. The reason why Tag 25 is upset is that it is waste gas recycled from another downstream unit to which the oscillation has propagated. The new non-linearity test shows that tag 25 is not the root cause because its time trend has no non-linearity. In fact, non-linearity analysis shows that Tag 25 is one of the of the secondary oscillations most far away from the root cause.

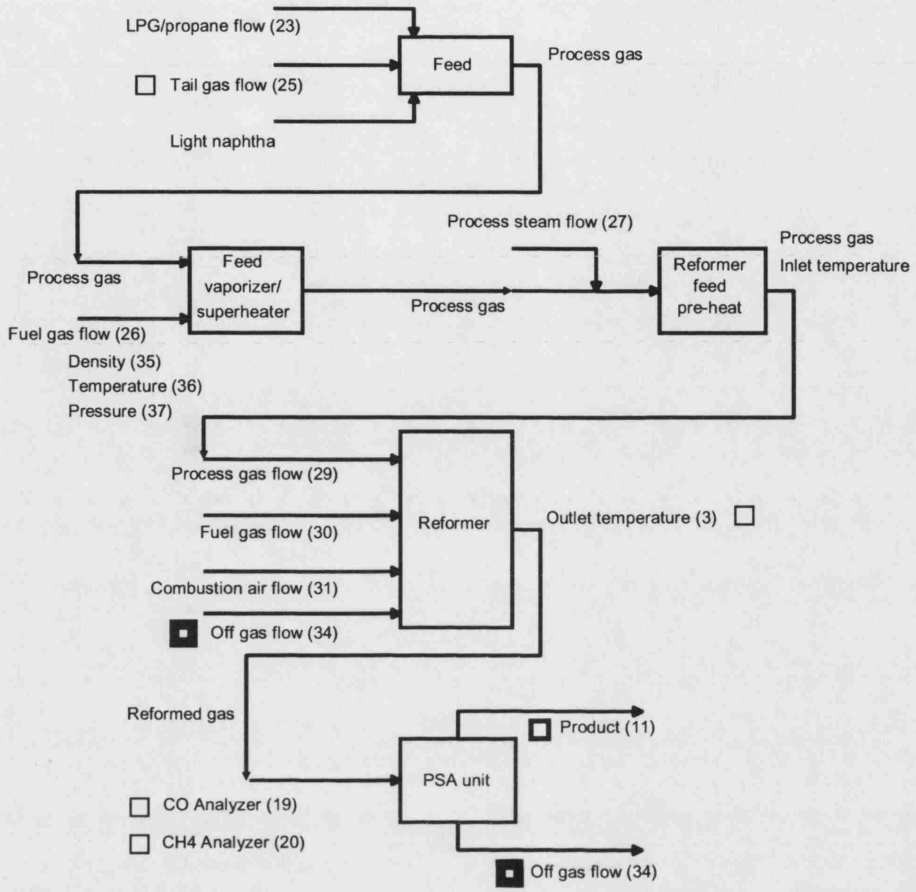

Fig 6.9 Process schematic of SE Asia hydrogen plant showing the location of the 16.7 minute oscillation (faint ☐ symbols) and highlighting non-linearity (heavy ☐ symbols). Non-linearity is most strong in tag 34.

6.4.2 Diagnosis of other root causes

Other root causes: Possible root causes besides non-linearity identified in section 6.2.1 include:

- Control loop interactions arising when two controllers have a shared mass and/or energy store (e.g. pressure and level controllers may compete for control of the contents of a reactor);
- Structural disturbances caused by coordinated transfers of mass and/or energy between different process units, especially when a recycle is present;
- Disturbances entering at plant boundaries (e.g. due to site utilities such as steam);
- Poorly tuned controllers.

There still remain open areas of research, but some suggestion for ways forward are presented in this sub-section.

Directions for future work: Normal-mode theory (Pain, 1987) shows that interactions happen when coupled systems share mass or energy storage and that they appear in the time trends of the measurement as beating or amplitude modulation. There is a good example of controller interaction with a beating pattern in a case study from Mitsui Chemicals (Matsuo *et. al.*, 2004) which suggests that normal-mode analysis will be successful. The frequencies and damping factors of the normal modes and their intensity at each measurement point would be determined from process measurements by Prony analysis, as is done for electricity transmission systems (Kundur, 1993). In Prony analysis the measured trends are decomposed as sums of normal modes having damped sinusoidal impulse responses. The key idea is that the measurements participating most strongly in the least damped modes are the closest to the shared mass or energy store responsible for the interaction This hypothesis has now to be examined critically and backed up by physical modelling.

Plants with recycles may have structural disturbances particularly when the inventory control or mass coupling is tight (Belanger and Luyben, 1997). It is speculated that mode frequencies and damping factors can also be utilized to detect a structural root cause, but curing it (e.g. to determine from plant data which of several level controllers should be de-tuned) will be a challenge. It seems likely that the diagnosis of structural disturbances will have to extend beyond the data-driven approach and it may need connectivity information from the process schematic to be incorporated into the analysis.

From observation, disturbances caused by poor controller tuning are often localized and do not propagate plant-wide. The validity and reasons for the observation have now to be elucidated by consideration of process frequency responses and mode damping. Examination of industrial examples tend to suggest that problems due to poor controller tuning mainly have an impact when they lead to other problems already discussed such as saturation of a manipulated variable and will therefore be already covered by the work presented in this chapter.

6.5 CHAPTER SUMMARY

The chapter proposed a method for locating the root cause of a distributed disturbance in a chemical process when the root cause is a non-linearity such as a sticking valve or process non-linearity. Such non-linearities cause persistent oscillations known as limit cycles, and can propagate around the plant causing numerous secondary disturbances. The key feature of the proposed method is a measure of non-linearity in the time series that is strongest closest to the root cause.

A non-linear time series is a sequence of measurements that cannot be modelled as the output of a linear system driven by random noise. Following the work of Kantz and Schreiber (1997) and others, the extent of non-linearity is determined by comparing the predictability of the suspected non-linear time series to the predictability of its surrogates. These surrogates are linearized versions of the same time series without the phase coherency that characterizes a non-linear time series and which makes it predictable. A non-linearity index was introduced which is large if the time series under test is much more predictable than the surrogates.

The contributions of the chapter are two-fold. Firstly, a key insight has been that non-linearity in the time trends of plant measurements is greatest in those measurements closest to the root cause because mechanical filtering by the plant makes the signals more linear as the disturbance propagates away. This enables the root cause to be distinguished from the secondary upsets. The second contribution is the optimization of the non-linearity detection algorithm for oscillating time trends so that it can be applied automatically. The parameters of the algorithm have been studied extensively and robust default settings determined which make it possible to roll out the algorithm into new applications with a minimum of customization.

The source of the main oscillatory disturbance in the SE Asia refinery data was determined using the method and the root cause that caused the plant-wide upset was located at Tag 34. Tag 25 is also oscillating but is clearly shown not to be the root cause by the non-linearity test even though it is the most upstream measurement in the plant.

7. STICTION IN CONTROL VALVES

7.1 OVERVIEW

7.1.1 Motivation

Industrial surveys by Bialkowski, (1992), Ender, (993) and Desborough and Miller, (2002) have shown that many distributed disturbances in chemical processes are caused by sticking control valves. If the control valve has a nonlinearity such as stiction or a deadband then a self-sustained limit cycle can arise in closed loop operation. The previous chapters showed how it is possible to detect disturbances and to diagnose those due to non-linear root causes such as sticking valves. The purpose of this chapter is to understand the reasons why a sticking control valve can have such a severe impact on a process through a detailed understanding of friction in control valves and of the mechanisms underlying limit cycle oscillations.

The aims of this chapter are:

- A new definition of the term *stiction* using terms similar to those used in the American National Standard Institution's (ANSI) formal definition of deadband;
- A physics-based model of a sticking valve that qualitatively matches the behaviour seen in real processes;
- A describing function analysis of the stiction model which gives insights into the nature of control loop limit cycles;
- Insight into a heuristic test used by industrial process control engineers to prove that a control loop has a sticking valve.

7.1.2 Introduction

This chapter analyses the effects of friction in control valves by means of first principles physical modelling and a describing function analysis. The results match well with those seen in industrial case studies and give insight into what to do to alleviate the problems caused by a control valve that has excessive friction.

The next section examines the characteristics of a feedback control loop with a sticking valve and gives some examples. It gives a new and precise definition for the term *stiction* which arose in a collaboration with Choudhury and Shah at the University of Alberta. A feature of the definition is

that it uses the same terminology and concepts as in ANSI definitions of deadband such that the effect is described in term of the input-output behaviour and not in terms of underlying physical mechanisms.

Section 7.3 discusses the physical mechanism leading to friction in control valves and gives a physical model based upon earlier work by Kayihan and Doyle (2000). It also describes a data-driven empirical model of a stiction non-linearity devised by Choudhury at the University of Alberta. The empirical model uses an engineering approach to the quantification of deadband by expressing it as a percentage of range of the valve demand signal.

The last section derives the describing function of a system with deadband and stiction. The describing function is a pseudo gain term for a non-linearity which depends on the amplitude of the driving signal. In closed loop, it can be used to give an approximate analysis of the period and magnitude of a limit cycle and also to determine the conditions under which a spontaneous limit cycle will arise. The chapter ends with some insights into the nature of limit cycles caused by valve deadband and stiction.

7.1.3 Work done by the thesis author and others

Work reported in this chapter has been published in the following paper:

> Choudhury, M.A.A.S., Thornhill, N.F., and Shah, S.L., 2005, Modelling of valve stiction, *Control Engineering Practice,* 13, 641-658.

The principal author was Shoukat Choudhury of the University of Alberta who wrote most sections of the paper. The section of the paper about physics-based modelling of a control valve was written by Thornhill and is the basis of section 7.2 in this chapter of the thesis. The new definition of stiction in section 7.2 was worked on by all authors of the paper. Theoretical work on describing function analysis in section 7.3 was a joint effort between Thornhill and Choudhury while the empirical model that is described in section 7.3 is the work of Shoukat Choudhury and Sirish Shah.

7.2 VALVE STICTION

7.2.1 Introduction

The challenge for this section of the thesis is to understand and describe the behaviour of a sticking control valve and to reproduce it by means of mathematical modelling. There is a particular focus on the behaviour denoted as *stiction* because the literature and existing Standards do not adequately explain or clearly describe a characteristic jumping effect observed in a system with stiction characteristics. A new definition is given of stiction in a control valve, devised in collaboration with Choudhury and Shah at the University of Alberta. It is cast in terms of the observed input-output behaviour rather than by reference to the physical mechanisms that can cause stiction.

7.2.2 Non-linear characteristics and stiction

Input-output behaviour of non-linear components: Figure 7.1. presents the idealized input-output behaviour of several types of non-linear systems and also shows the time trend of a deadband and stiction non-linearity being driven by a sinusoidal input. Those most likely to be found in a control valve are deadband and stiction illustrated in panels (a), (e) and (f).

Figure 7.2 shows the input-output characteristics of two industrial control loops showing deadband and stiction behaviour that were first presented in Thornhill and Hägglund (1997). These were cascade slave loops and the input was therefore a time-varying set point and the output a measurement of a flow. These plots demonstrate that the theoretical characteristics presented in Figure 7.1 may be observed in a practical situation.

Formal definitions: The following definitions of non-linearities are given by ANSI (ISA-S51.1-1979, Process Instrumentation Terminology):

> Deadband: In process instrumentation, it is the range through which an input signal may be varied, upon reversal of direction, without initiating an observable change in output signal. There are separate and distinct input–output relationships for increasing and decreasing signals. Deadband produces phase lag between input and output. Deadband is usually expressed in percent of (input) span.
>
> Hysteresis: It is that property of the element evidenced by the dependence of the value of the output, for a given excursion of the input, upon the history of prior excursions and the direction of the current traverse. It is usually determined by subtracting the value of deadband from the maximum measured separation between upscale-going and downscale-going indications of the measured variable (during a full-range traverse, unless otherwise specified) after transients have decayed. Some reversal of output may be expected for any small reversal of input. This distinguishes hysteresis from deadband.

> Dead zone: It is a predetermined range of input through which the output remains unchanged, irrespective of the direction of change of the input signal. There is but one input–output relationship. Dead zone produces no phase lag between input and output.
>
> Backlash: In process instrumentation, it is a relative movement between interacting mechanical parts, resulting from looseness, when the motion is reversed.
>
> Stiction: It is the resistance to the start of motion, usually measured as the difference between the driving values required to overcome static friction upscale and downscale.

ANSI is quite clear that deadband and hysteresis are not the same effect. The definition of deadband focuses on the lack of response to a changing input and is quantified in terms of input signal span. Hysteresis, by contrast, is the non-reproducibility of the output response and is quantified by the separation of the upgoing and downgoing responses on the y-axis. The above definitions also show that the term *backlash* specifically applies to the slack or looseness of the mechanical parts. Backlash can add deadband effects if there is slack in a rack-and-pinion type of actuator (Fisher-Rosemount, 1999) or if there is a loose connection in rotary valve shaft.

A new definition for stiction: The above ANSI definition of *stiction* was first proposed in 1963 in American National Standard C85.1-1963, Terminology for Automatic Control, and was not updated either in the 1979 or 1993 revisions. ANSI refers to the driving force or input, i.e. to the horizontal part of the little triangles appearing at the bottom left and top right in figs 7.1(e) and (f) but does not adequately capture the jump behaviour of the vertical part of the triangles which is commonly seen in sticking control valves. The characteristic vertical jump can also be seen in the time trends in the bottom panel of Figure 7.1. When the valve starts to move it jumps because the moving friction is less than the static friction.

Many authors have considered the issue of stiction (Armstrong- Hélloury *et. al.*, 1994; Aubrun *et. al.*, 1995; McMillan, 1995; Sharif and Grosvenor, 1998; Horch *et. al.*, 2000; Horch, 2000; Ruel, 2000; Gerry and Ruel, 2001). EnTech (1998) described stiction as a tendency to stick-slip due to high static friction and defined *stick-slip* as the tendency of a control valve to stick while at rest and to suddenly slip after force has been applied. Definitions proposed by these authors have tended to focus on the mechanics of stiction whereas the ANSI philosophy is to focus on the input-output behaviour. It was therefore argued in Choudhury *et. al.* (2005) that a new definition of stiction is needed and the following was proposed:

> Stiction: It is a property of an element such that its smooth movement in response to a varying input is preceded by a sudden abrupt jump called the slip-jump. Slip-jump is expressed as a percentage of the output span. Its origin in a mechanical system is static friction which exceeds the friction during smooth movement.

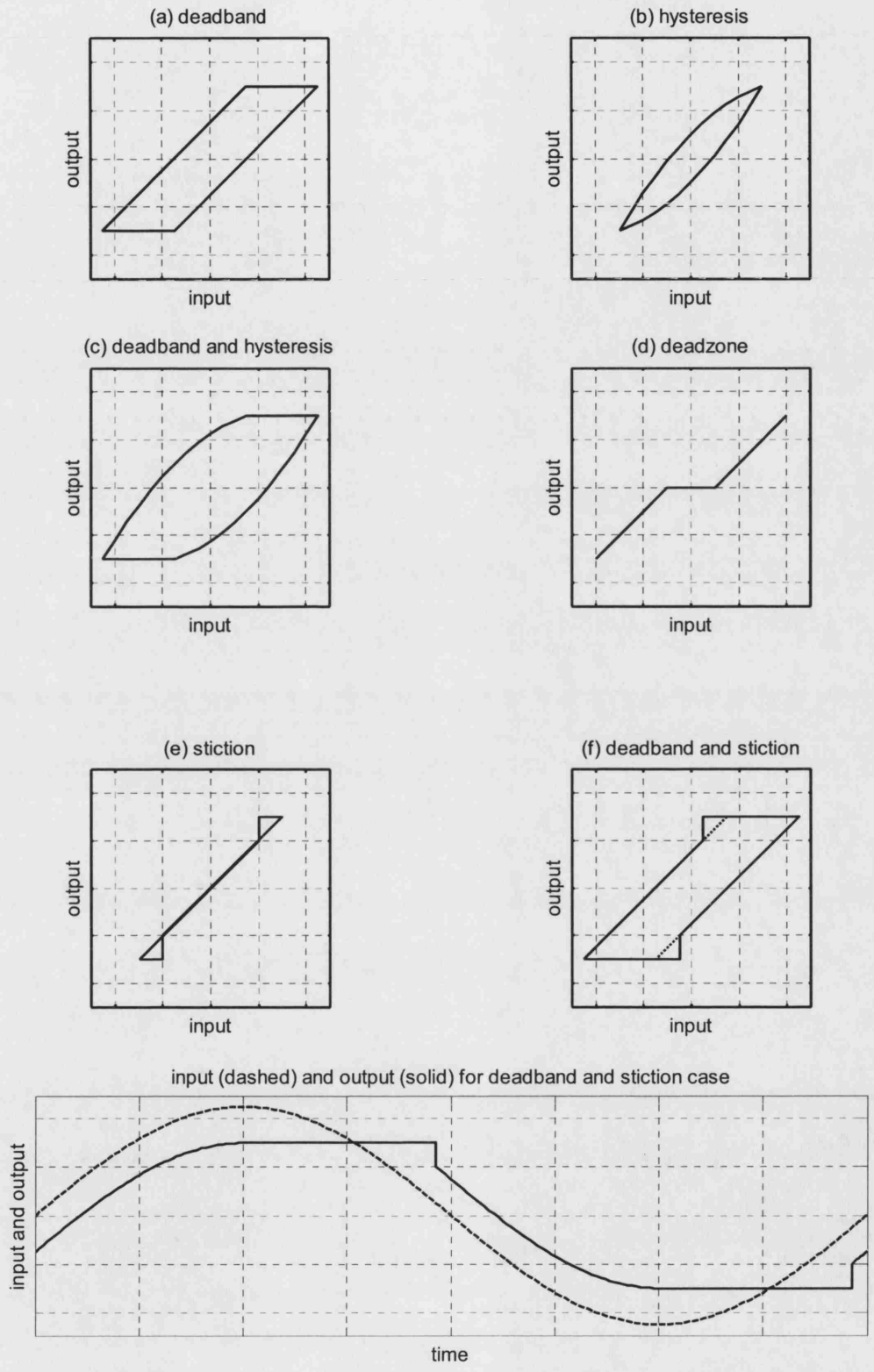

Fig 7.1 Upper panels: Characteristics of common non-linearities. Bottom panel: Response to a sinusoidal input of a deadband and stiction non-linearity.

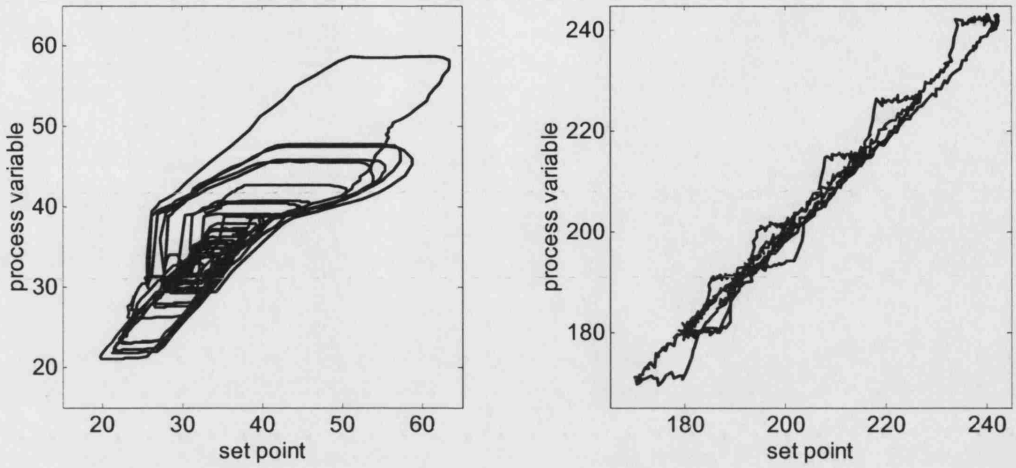

Fig 7.2 Industrial examples of control loops with sticking valves (from Thornhill and Hägglund, 1997). Left panel: Deadband Right panel: stiction.

7.3 A FIRST PRINCIPLES MODEL OF A STICKING VALVE

7.3.1 Control valves

The construction of a control valve: Figure 7.3 is a drawing showing the parts of a control valve. It was downloaded in 2002 from the Water Software Systems (WSS) group of the Faculty of Computing Sciences and Engineering at De Montfort University and its use here is gratefully acknowledged.

The aim is to adjust the valve so that it restricts the flow of process fluid through the pipe which can be seen at the very bottom of the figure. The valve plug (marked with an × symbol) is rigidly attached to a stem which is attached to a diaphragm in an air pressure chamber in the actuator section at the top of the valve. When compressed air is applied the diaphragm moves down and the valve closes. The spring is compressed at the same time. When the air pressure is reduced the spring forces the diaphragm back up and the valve opens. Therefore the valve in Figure 7.3 is a fail-open type of valve.

Friction in a control valve: The friction in the valve arises principally in the packing. It is the packing that stops process fluid from leaking out of the valve but the valve stem nevertheless has to move freely relative to the packing. There is therefore a trade-off because too-tight packing

reduces emissions and leaks from the valve but at the same time increases the friction. Loose packing reduces friction but there will be more leakage of process fluids. Other effects that cause excessive friction are corrosion of the valve stem which makes it rough, and deposits on the valve seat which can make the valve plug stick in the seat.

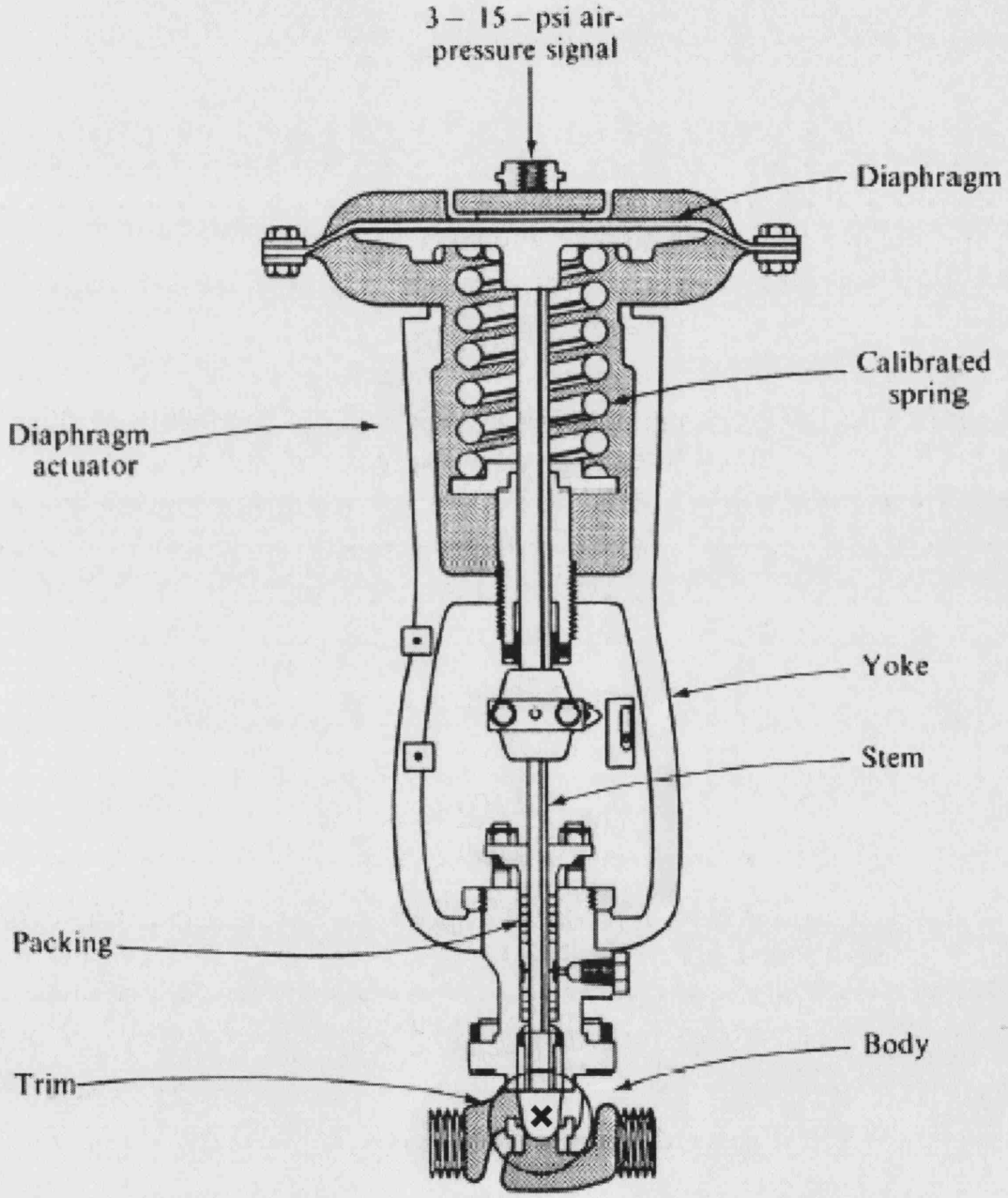

Fig 7.3 Drawing of a control valve. The point marked ✖ indicates the valve plug. Downloaded (2002) from http://www.eng.dmu.ac.uk/wssys/staff/IndCont/lect04.pdf

7.3.2 Physical modelling of a control valve

The physics of a control valve: For a pneumatic sliding stem valve, the force balance equation based on Newton's second law (Fitzgerald, 1995; Kayihan and Doyle, 2000; Whalen, 1983), can be written as:

$$M\frac{d^2x}{dt^2} = \sum Forces = F_a + F_r + F_f + F_p + F_i$$

where:

- M is the mass of the moving parts
- x is the stem position
- F_a the force applied by the pneumatic actuator $F_a = Au$ where A is the area of the diaphragm and u is the actuator air pressure or the valve input signal
- $F_r = -kx$ is the spring force where k is the spring constant
- $F_p = -\alpha\Delta P$ is the force due to ΔP, the fluid pressure drop across the valve
- F_i is an extra force required to force the valve to be in the seat
- F_f is the friction force

Following Kayihan and Doyle (2000) F_i and F_p are assumed to be zero because of their negligible contribution in the model.

Friction model: The friction model is from Karnopp (1995), Olsson (1996) and was used also by Horch (1999). It includes static and moving friction. The moving friction is the first line of the friction expression below and comprises a velocity independent term F_c known as Coulomb friction and a viscous friction term vF_v that depends linearly upon velocity. Both act in opposition to the velocity, as shown by the negative signs:

$$F_f = \begin{cases} -F_c\,\mathrm{sgn}(v) - vF_v & \text{if } v \neq 0 \\ -(F_a + F_r) & \text{if } v = 0 \text{ and } |F_a + F_r| \leq F_s \\ -F_s\,\mathrm{sgn}(F_a + F_r) & \text{if } v = 0 \text{ and } |F_a + F_r| > F_s \end{cases}$$

The second line in the above equations applies when the valve is stuck. The velocity of the stuck valve is zero and not changing, therefore the acceleration is zero also. Thus the right hand side of Newton's law is zero and $F_f = -(F_a + F_r)$. The third line of the model represents the situation at the instant of breakaway. At that instant the sum of forces is $(F_a + F_r) - F_s\,\mathrm{sgn}(F_a + F_r)$ which is not zero if $|F_a + F_r| > F_s$, where F_s is the maximum static friction. Therefore the acceleration becomes non-zero and the valve starts to move.

Model parameters: The physical model of a control valve requires several parameters to be known. The mass M and typical friction forces depend upon the design of the valve. Kayihan and Doyle (2000) used manufacturer's values suggested by Fitzgerald, (1995) and similar values

have been chosen here apart from a slightly increased value of F_s and a smaller value for F_c in order to make the demonstration of the slip-jump more obvious. The values have also been converted to SI units. Figure 7.4 shows the friction force characteristic attained with these values. The friction force opposes the velocity hence it is negative when the velocity is positive.

The calibration factor of Table 7.1 is required because the required stem position x_r is the input to the simulation. In the absence of stiction effects the valve moving parts would come to rest when the force due to air pressure on the diaphragm is balanced by the spring force. Thus at rest $Au = kx$ and so the calibration factor relating air pressure u to x_r is k/A. The consequences of miscalibration are discussed below.

	Kayihan and Doyle III, 2000	**Nominal case**
M	3 lb (1.36 kg)	1.36 kg
F_s	384 lbf (1708 N)	1750 N
F_c	2320 lbf (1423 N)	1250 N
F_v	3.5 $\text{lbf}\cdot\text{s}\cdot\text{in}^{-1}$ (612 $\text{N}\cdot\text{s}\cdot\text{m}^{-1}$)	612 $\text{N}\cdot\text{s}\cdot\text{m}^{-1}$
spring constant, k	300 $\text{lbf}\cdot\text{in}^{-1}$ (52500 $\text{N}\cdot\text{m}^{-1}$)	52500 $\text{N}\cdot\text{m}^{-1}$
diaphragm area, A	100 in^2 (0.0645 m^2)	0.0645 m^2
calibration factor	none	807692 $\text{Pa}\cdot\text{m}^{-1}$
air pressure (typ)	10 psi (68950 Pa)	68950 Pa
valve travel (typ)	3 in (0.076 m)	0.076 m
velocity (typ)	5 $\text{in}\cdot\text{s}^{-1}$ (0.127 $\text{m}\cdot\text{s}^{-1}$)	0.127 $\text{m}\cdot\text{s}^{-1}$

Table 7.1 Nominal values used for physical valve model.

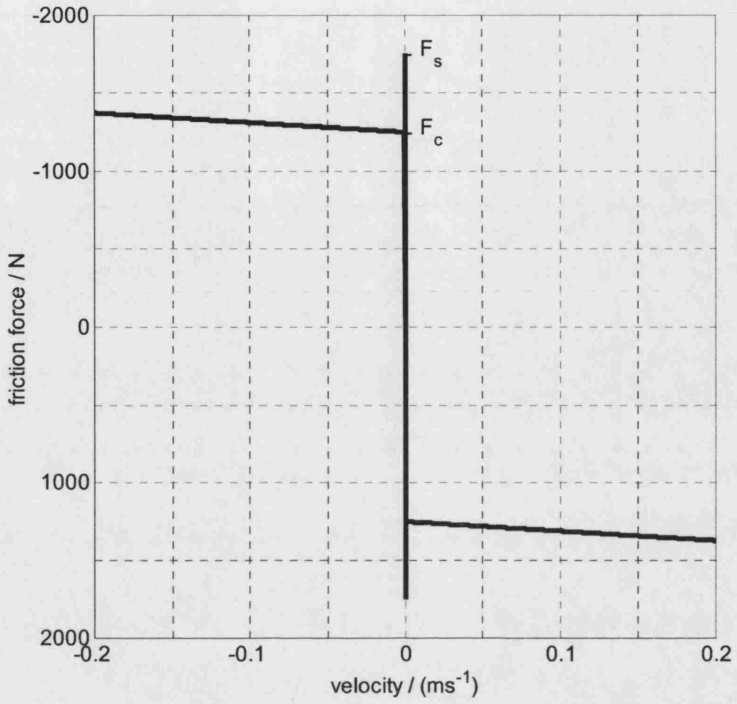

Fig 7.4 Friction model of a valve.

Detection of zero velocity: A numerical simulation of the valve requires a condition for the velocity to be zero in order for the second and third lines of the friction model to be activated. The velocity in $\mathrm{m\,s^{-1}}$ is taken to be zero if it falls in the range $-0.0005 \le v \le 0.0005$.

Model of the pressure chamber: It was pointed out by Kano (personal communication) that the above model does not respond correctly to a stochastic input in which the air pressure u changes randomly at each simulation step. The reason is that the zero velocity condition is not met unless the amplitude of the stochastic input is small.

A physical valve would not experience a stochastic air pressure signal because the air pressure chamber has a smoothing effect. A solution to the numerical problem identified by Kano can make use of this physical insight by introducing a first order lag transfer function model to represent the dynamics of the pressure chamber in place of the algebraic model $F_a = Au$. The adjusted model is:

$$F_a = \frac{A}{\tau s + 1} u$$

where τ is a time constant which can be set to some realistic value such as 1 second.

If the stochastic input u has a large enough amplitude then the valve does not stick even when the pressure chamber dynamics are taken into account. This effect is also observed with a real

valve and such an input is called a dither signal. Making random changes to the valve air pressure is well known as a means of reducing the impact of a control valve deadband.

7.3.3 Results from simulation

Open loop response: Figure 7.5 shows the valve position when the physics based valve model is driven by a sinusoidal variation in *op* in open loop in the absence of the controller. The left hand column shows the time trends and the right hand panels are plots of valve demand (*op*) versus valve position (*mv*). Several cases are simulated using the parameters shown in Table 7.2. The "linear" values are those suggested by Kayihan and Doyle (2000) for the best case of a smart valve with Teflon packing.

	linear	deadband only	slip-jump and deadband (open loop)	slip-jump and deadband (closed loop)	slip-jump only
F_S/N	45	1250	1750	1000	1750
F_C/N	45	1250	1250	400	0
$F_V/\left(\mathrm{N\cdot s\cdot m^{-1}}\right)$	612	612	612	612	612

Table 7.2 Friction values used in simulation of physical valve model.

In the first row of Figure 7.5, the Coulomb friction F_C and static friction F_S are small and linear viscous friction dominates. The input and output are almost in phase because the sinusoidal input is of low frequency compared to the bandwidth of the valve model and is on the part of the frequency response function where input and output are in phase.

Valve deadband is due to the presence of the Coulomb friction F_C, a constant friction term which acts in the opposite direction to the velocity. Deadband effects are seen in the presence of Coulomb friction even when there is no excessive static friction. In the deadband simulation case the static friction is the same as the Coulomb friction, $F_S = F_C$. The deadband arises because, on changing direction, the valve remains stationery until the net applied force is large enough to overcome F_C. The deadband becomes larger if F_C is larger.

A valve with high static friction such that $F_S > F_C$ exhibits jumping behaviour that is different from a deadband, although both behaviours may be present simultaneously. When the valve just starts to move the friction force reduces abruptly from F_S to F_C. There is therefore a discontinuity in the model on the right hand side of Newton's second law and a large increase in acceleration of the valve moving parts. The initial velocity is therefore faster than in the pure deadband case leading to the jump behaviour observed in the third row of Figure 7.5. If the Coulomb friction F_C is absent then the deadband is absent and the jump allows the *mv* to catch up with the *op* (fourth row).

If the valve is miscalibrated then swings in the valve position (*mv*) are larger than swings in the demanded position (*op*). In that case the gradient of the *op-mv* plot is greater than unity during the moving phase. The bottom row of Figure 7.5 shows the case when the calibration factor was too large by 25%. Stiction was also used in this simulation.

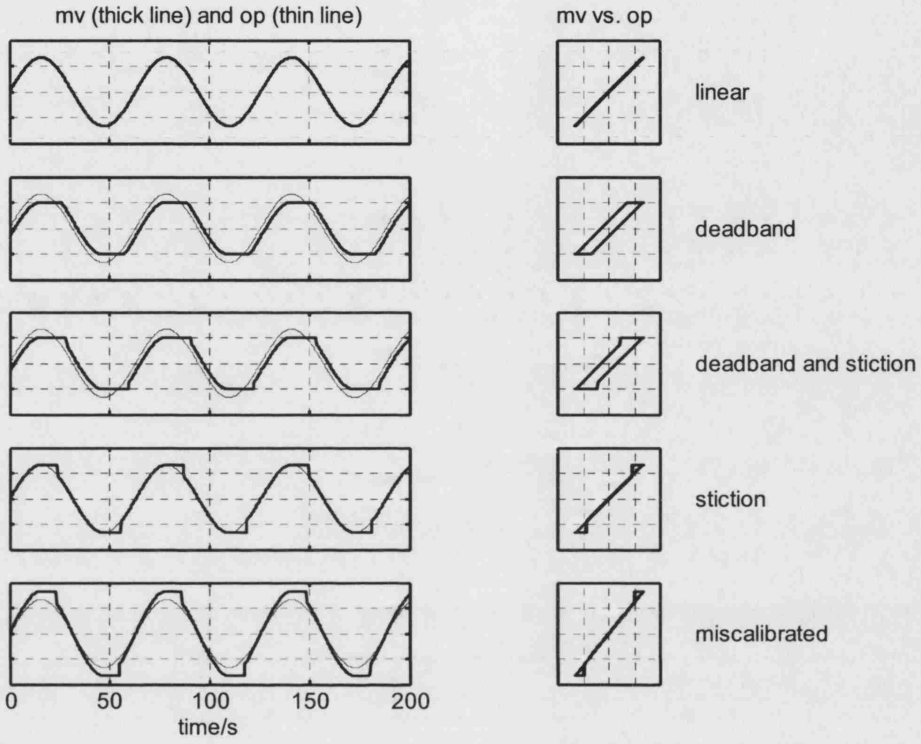

Fig 7.5 Open loop responses of physics based model.

Closed loop dynamics: In a simulation of closed loop behaviour the valve output drives a first order plus time delay (FOPTD) process $G(s)$ and receives its *op* reference input from a proportional plus integral (PI) controller:

$$G(s)=\frac{3\times e^{-10s}}{10s+1} \qquad C(s)=0.2\left(\frac{0.1s+1}{0.1s}\right)$$

Figure 7.6 shows limit cycles induced in this control loop by the valve together with the plots of valve demand (*op*) and valve position (*mv*). Limit cycles were present in some cases even though the set point to the loop was zero. That is, they were spontaneously generated and sustained by the loop in the absence of an input.

There was no limit cycle for the FOPTD process in the linear valve case dominated by viscous friction or in the case with deadband only, i.e. when $F_s = F_c$. It is known that deadband alone cannot induce a limit cycle unless the process $G(s)$ has integrating dynamics, as will be discussed shortly.

The presence of stiction ($F_s > F_c$) induces a limit cycle with a characteristic triangular shape in the controller output. Cycling occurs because an offset exists between the set point and the output of the control loop while the valve is stuck, which is integrated by the PI controller to form a ramp. By the time the valve finally moves in response to the controller *op* signal the actuator force has grown quite large and the valve moves quickly to a new position where it then sticks again. Thus a self limiting cycle is set up in the control loop.

If stiction and deadband are both present then the period of the limit cycle oscillation can become very long. The combination $F_s = 1750\,\text{N}$ and $F_c = 1250\,\text{N}$ gave a period of 300s while the combination $F_s = 1000\,\text{N}$ and $F_c = 400\,\text{N}$ had a period of about 140s (top row, Figure 7.6), much longer than the time constant of the controlled process or its cross-over frequency. The period of oscillation can also be influenced by altering the controller gain. If the gain is increased the linear ramps of the controller output signal are steeper, the actuator force moves through the deadband more quickly and the period of the limit cycle becomes shorter. The technique of changing the controller gain is used by industrial control engineers to test the hypothesis of a limit cycle induced by valve non-linearity while the plant is still running in closed loop.

In the stiction-only case shown in the third row of Figure 7.6 the Coulomb friction is negligible and the oscillation period is shorter because there is no deadband in the valve. The bottom row in Figure 7.6 shows that miscalibration and stiction cause an overshoot in closed loop.

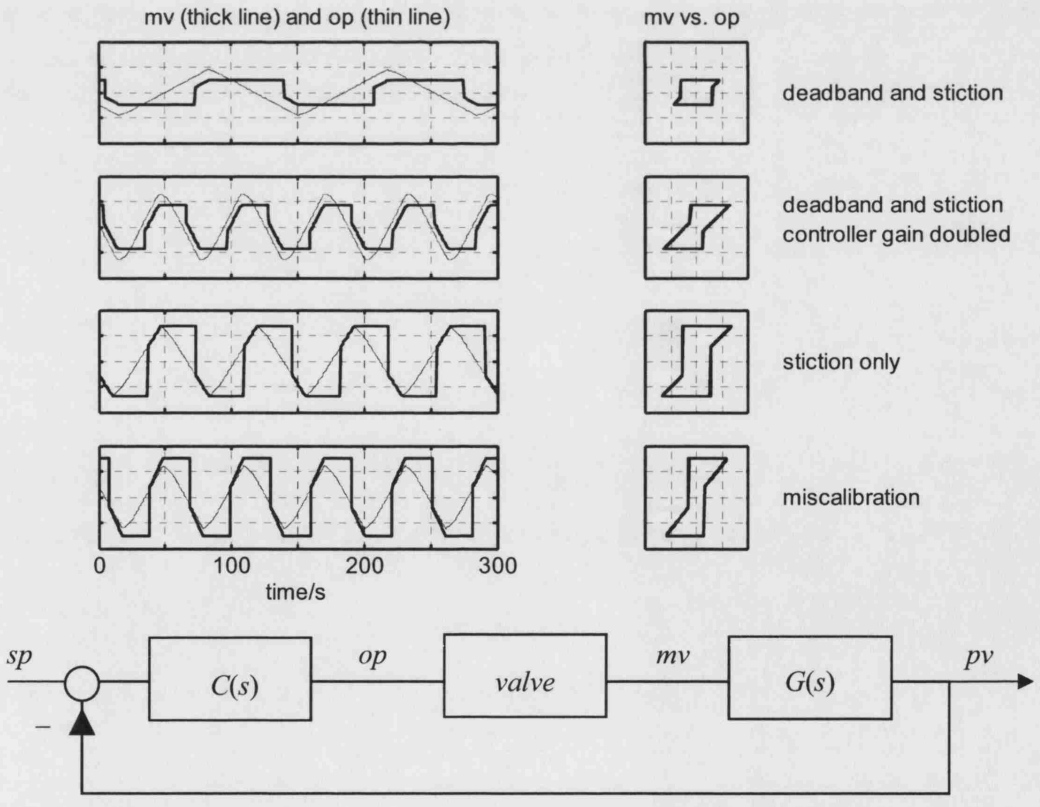

Fig 7.6 Lower panel: The closed loop configuration for simulation of first order plus deadtime process with PI controller in the presence of the physics based valve model. Upper panels: Closed loop responses.

7.3.4 An empirical model for valve stiction

The need for an empirical model: The advantage of a physical model is that the effects of changes in the parameters can be related to changes in behaviour through first principles. For instance, modelling shows that the physical origin of a deadband is a non-zero coulomb friction, $F_c \neq 0$.

However, the parameters of a physical model such as mass of the moving parts of the valve, spring constants, and forces are not normally known. Accurate values have to be determined from the manufacturer or by testing. Working with a physical model may therefore become time consuming and cumbersome. Also, in industrial practice stiction and deadband are identified as a percentage of the valve travel or span of the valve input signal. The empirical data-driven model of stiction was devised by Choudhury of the University of Alberta to overcome these difficulties.

An empirical valve model: The proposed data-driven model has parameters that can be directly related to plant data and it produces the same behaviour as the physical model. In the process industry, stiction is generally measured as a percentage of the valve travel or the span of the control signal (Gerry and Ruel, 2001). For example, a 2% stiction means that when the valve gets stuck it will start moving only after the cumulative change of its control signal is greater than or equal to 2%. If the range of the control signal is 4–20 mA, then a 2% stiction means that a change of the control signal less than 0:32 mA in magnitude will not be able to move the valve.

Figure 7.7 shows the flowchart detailing the rules and equations which comprise the empirical model. A full description can be found in Choudhury *et al.* (2005) where results are shown that closely match the physics based model. Further developments to Choudhury's basic model have been proposed by Kano *et. al.* (2004).

Key features of the model are:

- The models accommodates the 4-20 mA ranges of industrial instrumentation systems;
- The model needs only the *op* signal from the controller (on a 4-20 mA scale) and the specification of deadband and stickband as a percentage of range;
- Valve opening characteristics such as linear or equal percentage are included.

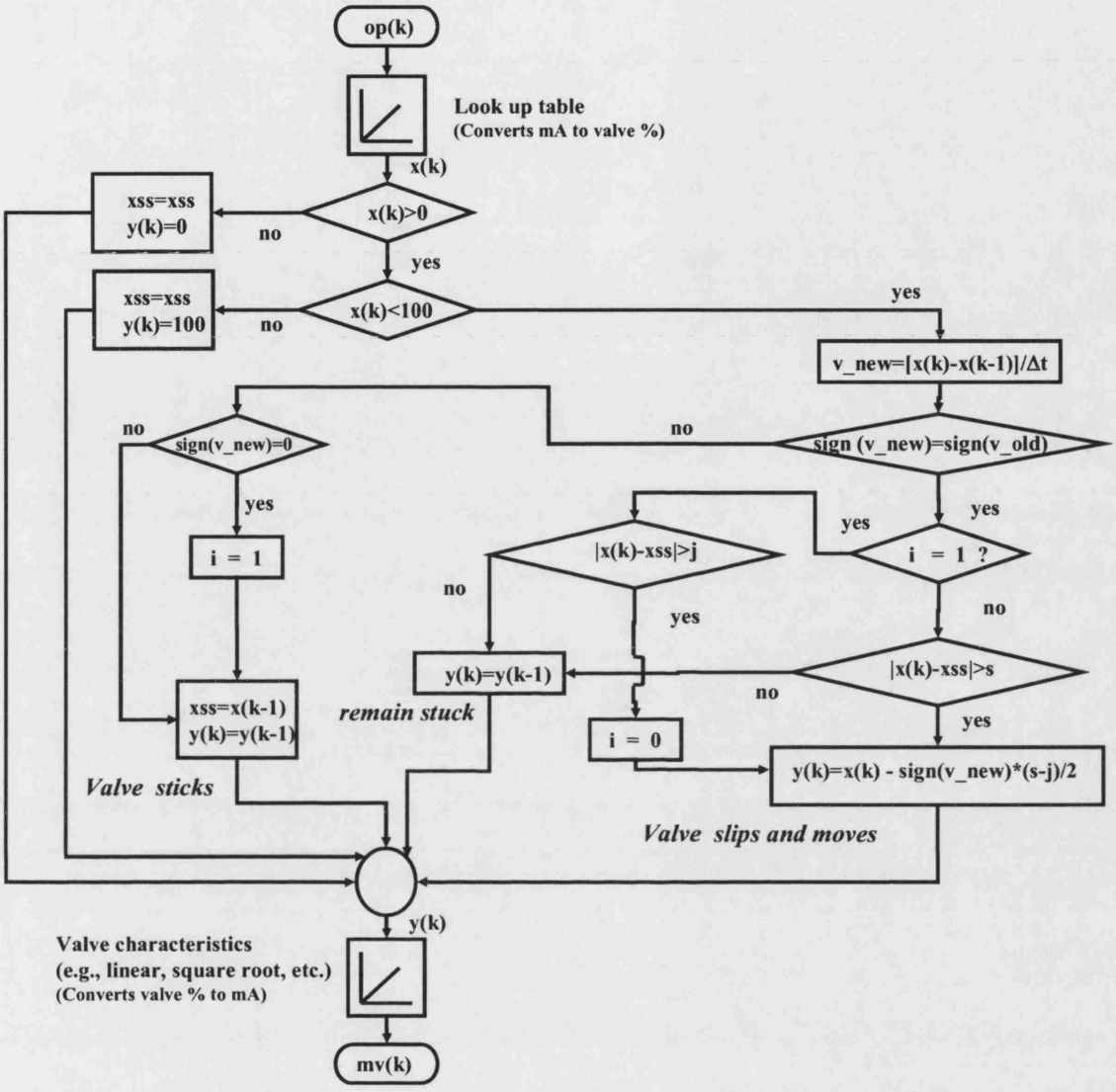

Fig 7.7 A data-driven model for valve stiction, provided courtesy of M.D. Shoukat Choudhury, University of Alberta.

7.4 DESCRIBING FUNCTION ANALYSIS

7.4.1 Describing function of stiction and deadband

Aims: The aims of describing function analysis (see e.g. Cook, 1986) are to gain insights into the simulation results and industrial observations presented in the paper. The non-linearities being considered are deadband and stiction, as shown in Figure 7.8. The bottom panel of Figure 7.1 shows the response of the non-linearity to a sinusoidal input.

The analysis is formulated in terms of the deadband, D, stickband, J and the sum of the two, $S = D + J$. Section 7.4.2 outlines the steps in the analysis and states the main results, while the details are presented in Appendix 3. Insights gained from describing function analysis are discussed and the analysis successfully predicted the behaviour of an industrial level control loop at the Eastman Chemical Company.

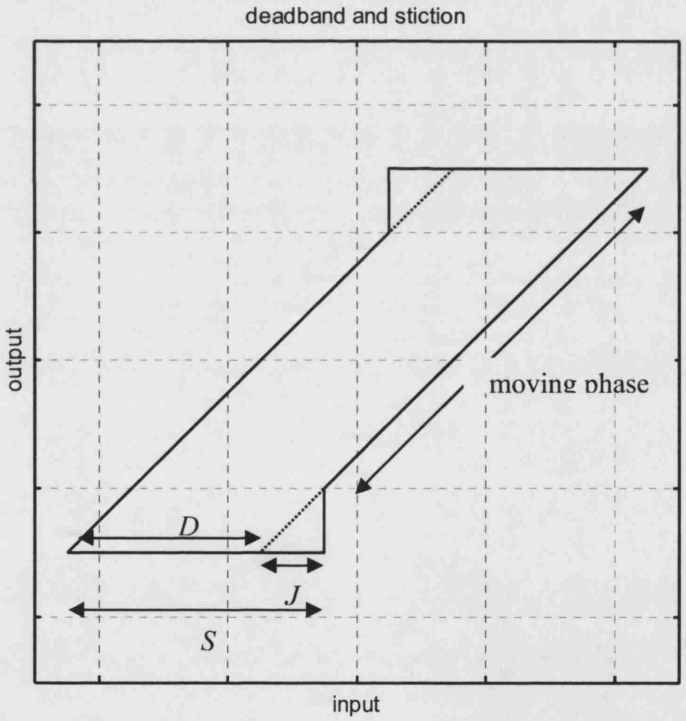

Fig 7.8 The input-output characteristics of a deadband with stiction non-linearity showing deadband D, stickband J and the sum of the two, $S = D + J$.

Background: In describing function analysis the non-linearity is modelled by a non-linear gain N The assumptions inherent in the approximation are that there are periodic signals present in the system and that the controlled system is low pass and responds principally to the fundamental

Fourier component. The conditions for oscillation in a negative feedback loop arise when the loop gain is -1:

$$G_o(i\omega) = -\frac{1}{N(X_m)}$$

where $G_o(i\omega)$ is the open loop frequency response which includes the controlled system and the proportional plus integral controller, and $N(X_m)$ is the describing function which depends on the magnitude of the controller output X_m. When the condition $G_o(i\omega) = -1/N(X_m)$ is met the system will spontaneously oscillate with a limit cycle. The variation of the quantity $-1/N(X_m)$ with signal amplitude means that signals initially present in the loop as noise can grow until they are big enough to satisfy the equality and hence provide a self-starting oscillation. The solution to the complex equation $G_o(i\omega) = -1/N(X_m)$, if one exists, may be found graphically by superposing plots of $G_o(i\omega)$ and $-1/N$ on the same set of axes.

An expression for the describing function: The describing function of a non-linearity is:

$$N = \frac{Y_f}{X}$$

where X is a harmonic input to the non-linearity of angular frequency ω_o and Y_f is the fundamental Fourier component angular frequency ω_o of the output from the non-linearity. Thus a Fourier analysis is needed on the output signal shown as a bold line in the bottom panel of Figure 7.1.

The quantity N depends upon the magnitude of the input X_m. N is complex for the stiction non-linearity because the output waveform has a phase lag compared to the input. The describing function is derived in Appendix 3 where it is shown that:

$$N = -\frac{1}{\pi X_m}(A - iB)$$

where

$$A = \frac{X_m}{2}\sin 2\phi - 2X_m \cos\phi - X_m\left(\frac{\pi}{2} + \phi\right) + 2D\cos\phi$$

$$B = -3\frac{X_m}{2} + \frac{X_m}{2}\cos 2\phi + 2X_m \sin\phi - 2D\sin\phi$$

$$\phi = \sin^{-1}\left(\frac{X_m - S}{X_m}\right)$$

The quantities S and D are defined in Figure 7.8.

7.4.2 Insights from describing function analysis

Asymptotes of the describing function: There is no output from the non-linearity if the peak-to-valley range of the input is less than S, i.e if $X_m < S/2$, because the moving phase only starts once the input signal moves beyond the flat part of the characteristic. Therefore the two extreme cases are when $X_m = S/2$ and $X_m \gg S$.

When $X_m \gg S$, then the effects of the deadband and slip-jump are negligible and the non-linearity in Figure 7.8 becomes approximately a straight line at 45° because the moving phase extends much further if the input signal has a large magnitude. The output is in phase with the input and $N = 1$. Thus $-1/N(X_m) = -1$ when $X_m \gg S$.

In the limit when $X_m \to S/2$ then the output is as shown in Figure 7.9. The left hand plot shows the output for a slip-jump with no deadband ($S = J$, $D = 0$) while the right hand plot shows a magnified plot of a dead-band with no slip jump ($S = D$, $J = 0$). In both cases the output lags the input by one quarter of a cycle in the limit when $X_m = S/2$. The output is a square wave of magnitude X_m in the $S = J$, $D = 0$ case and the describing function is $N = \frac{4}{\pi} e^{-i\pi/2}$. For the dead-band with no slip jump ($S = D$, $J = 0$) case, the output magnitude is very small. The describing function is $N = \varepsilon e^{-i\pi/2}$ where $\varepsilon \to 0$ as $X_m \to S/2$. The appendix provides detailed calculations of these results and also shows for the general case that the describing function limit when $X_m = S/2$ is:

$$N = \frac{4}{\pi} \times \frac{J}{S} e^{-i\pi/2}$$

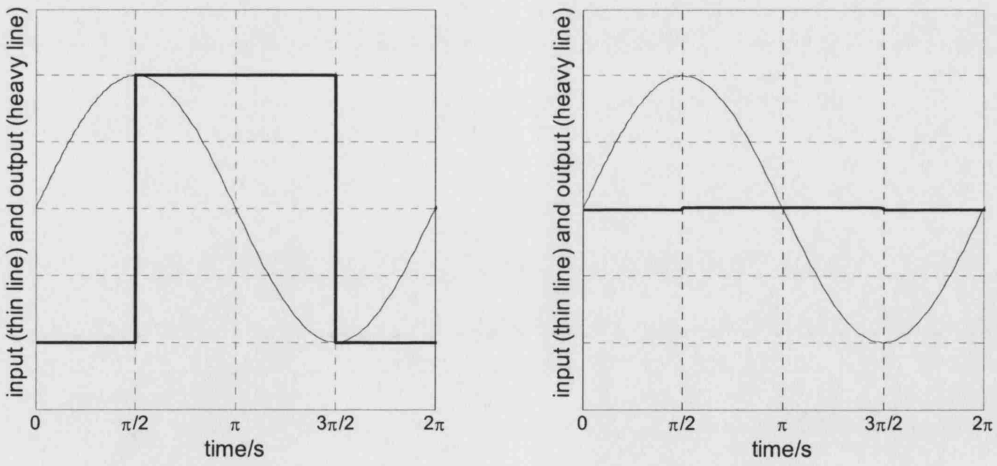

Fig 7.9 Input (thin line) and output (heavy line) time trends for the limiting case as $X_m = S/2$. Left panel: Slip-jump only with $S = J$. Right panel: Deadband only with $J = 0$. The output in the left plot has been magnified for visualisation; its amplitude becomes zero as X_m approaches $S/2$.

Insights gained from the describing function: Figure 7.10 shows graphical solutions to the limit cycle equation $G_o(j\omega) = -1/N(X_m)$ for the FOPTD process (left panel) and an integrating process (right panel), both with PI control. The describing function is parameterised by X_m and the open loop frequency response function of the controller and controlled system is parameterised by ω. Both systems are closed loop stable and thus intersect the negative real axis between 0 and -1.

Limit cycles of an FOPTD loop with PI control: The left hand panel of Figure 7.10 shows that the stiction forces the $-1/N$ curve onto the negative imaginary axis in the $X_m = S/2$ limit. Thus the frequency response curve of the FOPTD composition loop and its proportional plus integral controller is guaranteed to intersect with the describing function because the integral action means open loop phase is in the third quadrant of the complex plane at low frequency. Therefore:

- There will be a limit cycle for an FOPTD process under PI control if stiction is present (i.e. if $J \neq 0$).

The figure also shows the $-1/N$ curve for the deadband limit cycle ($J = 0$). In the $X_m = S/2$ limit the curve becomes large, negative and imaginary and the frequency response curve does not intersect with it. Therefore:

- The FOPTD process does not have a limit cycle if the non-linearity is a pure deadband, because the frequency response curve does not intersect the $-1/N$ curve.

The lack of a limit cycle in this case has been noted by other authors (Piipponen, 1996; McMillan, 1995). In fact a limit cycle with a deadband non-linearity *can* occur if the slope of the moving phase is greater than unity because the $-1/N$ curve moves to the right and, if it moves far enough, will intersect the frequency response curve. For instance, if the slope in the moving phase were 4 then the $X_m \gg S$ limit of the describing function would be $N = 4$ and the $-1/N$ curve would end at -0.25 on the complex plane and not at -1. However, this situation is unlikely to arise in a control valve where the slope of the moving phase is 1.

Limit cycles of an integrating loop with PI control: An integrating process with proportional plus integral control has a frequency response for which the phase becomes $-\pi$ at low frequency. The right hand panel of Figure 7.10 shows that it will intersect the $-1/N$ curves both for the stiction cases ($J \neq 0$) and also for the pure deadband case ($J = 0$). Therefore:

- A valve with a deadband and no stiction can cause a limit cycle oscillation for an integrating process with a PI controller.

The frequency of oscillation is higher and the period of oscillation shorter when the stiction is present because the $-1/N$ curves with the stiction intersect the frequency response curve at higher frequencies than the $-1/N$ curve for the deadband. Process control engineers will often switch an oscillating level or pressure controller to P-only in an attempt to stop limit cycling. This action changes the low frequency behaviour of the frequency response function to be more

like that of the loop with FOPTD plus PI controller (i.e. only one integrator) so that it no longer intersects the $-1/N$ curve for the deadband.

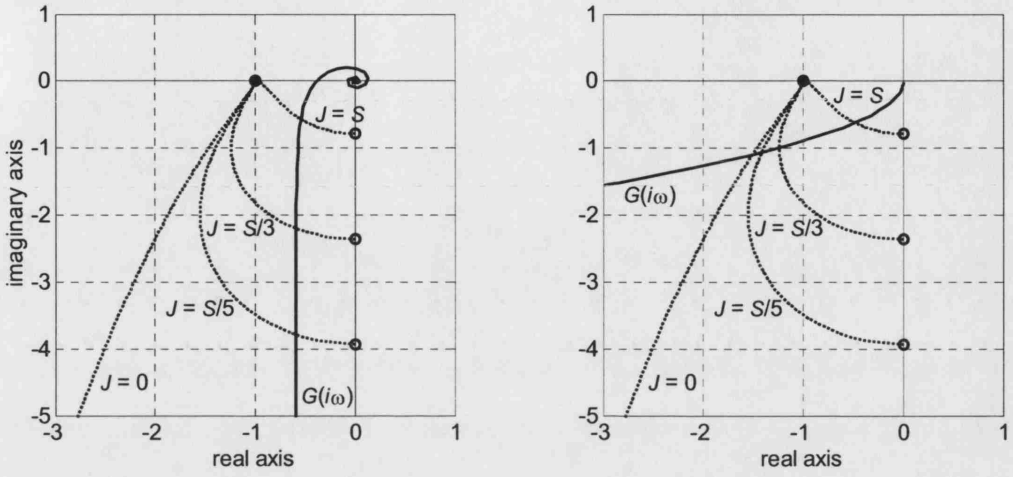

Fig 7.10 Graphical solutions for limit cycle oscillations. Left panel: FOPTD control loop. Right panel: integrating control loop. Dashed lines are the $-1/N$ curves and the solid line is the frequency response function.

Comparison with a real system: An analysis in Chapter 8 of a plant-wide disturbance at an Eastman Chemicals plant will identify a particular level control valve as being a candidate for the root cause. A valve with excessive friction was suspected.

The Eastman engineers checked for the presence of valve friction by doubling the controller gain. Figure 7.11 shows the plant test data from Eastman and a simulation of doubling of the gain using the first principles physics-based model. The simulation showed similar waveforms as the industrial data and also a decrease in period and amplitude when the controller gain was doubled.

Waveform shapes: A feature of interest in the test and simulation is the shapes of the curves which are different from those of Figure 7.6 for the FOPTD process where *op* was a triangular waveform. In Figure 7.11 the *pv* is a triangular waveform and *op* has a curved shape. The explanation is that the process is an integrator.

In both Figure 7.6 and 7.11 the *mv* (the flow through the sticking valve) has a square-shaped waveform because of its stuck phase. In Figure 7.6 for the FOPTD system, the *pv* (not shown) also has a square-shaped waveform similar to that of the *mv*. The square-shaped *pv* was integrated by the PI controller in the to become a ramp-shaped *op*.

In Figure 7.11, by contrast, the controlled system $G(s)$ was an integrator and the controller $C(s)$ also had an integrator. Therefore the *pv* is ramp-like because it is an integrated square waveform. The *op* is a series of parabolic segments since the result of integration of a ramp is a parabola.

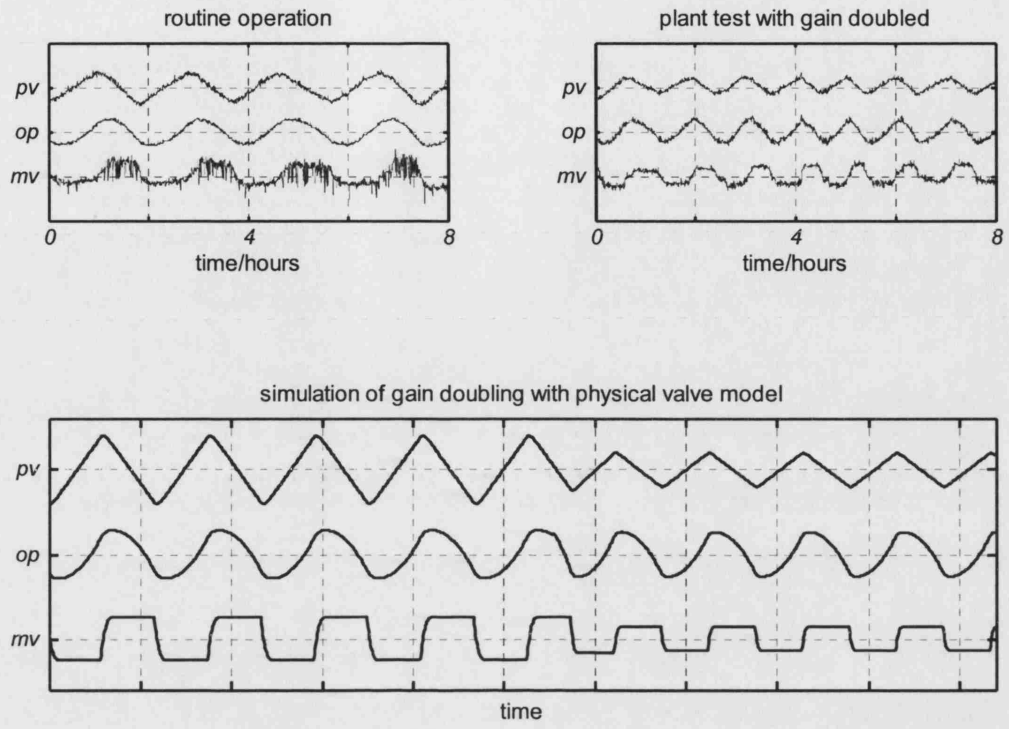

Fig 7.11 Response to a doubling of controller gain of a level control loop with a sticking valve. Upper panels: Plant tests at Eastman Chemical Company, Lower panel: Results from simulation. The meanings of the signals *pv*, *op* and *mv* are indicated in Fig 7.6.

Insights into the real system from the describing function: The qualitative match in Figure 7.11 is good. Both the plant test and the simulation show a reduction in period and amplitude when the gain was doubled. Describing function analysis can explain these findings.

The controlled system was an integrator under PI control and therefore the right hand plot of Figure 7.10 is relevant to the discussion below. It shows that both deadband-only ($J=0$) and deadband with stiction are expected to give a limit cycle in closed loop. The effect on the plot of $G(i\omega)$ curve of doubling the controller gain is that every point on the $G(i\omega)$ curve moves twice as far from the origin. The origin is the high frequency end of the $G(i\omega)$ curve so all the intersections with the $-1/N$ curves happen at higher frequency when the controller gain doubles. The conclusion is that the frequency of the limit cycle will increase (i.e. the period of oscillation decreases) when the controller gain doubles, as is seen in Figure 7.11.

7.5 CHAPTER SUMMARY

Chapter 7 gave an analysis of the behaviour of feedback control loops with sticking valves through first principles physical modelling together with a describing function analysis. It also reported a suggested new definition for stiction that was devised with Choudhury and Shah at the University of Alberta. Stiction arises when the static friction is larger than the moving friction and is a different effect than a deadband. The new definition fits alongside existing ANSI definitions because it is expressed in terms of the input-output behaviour whereas other attempted definitions of stiction have invoked descriptions of the mechanism.

The first principles physical model successfully reproduced a range of closed loop limit cycles and gave an excellent qualitative match to a limit cycle observed in an industrial process. The simulation showed the same changes in oscillation period and amplitude as the real process when the controller gain was changed.

Further analytical insights were achieved through describing function analysis. The analysis was successful in showing when a limit cycle will arise and when it will not. It provided an explanation for observations reported in the literature that only integrating processes oscillate if the valve has a pure deadband. For non-integrating processes the valve must also have stiction before a limit cycle can arise.

The physical model has fulfilled its purpose of giving in-depth understanding of the origin of deadband and stiction effects in a control valve. However, it is not useful for routine use because the values of the physical parameters such as friction forces and spring constant are not generally known. A data-driven empirical model devised by Choudhury at the University of Alberta was discussed as a means of overcoming this disadvantage.

8. INDUSTRIAL CASE STUDY

8.1 OVERVIEW

8.1.1 Motivation

This chapter presents a case study which uses the techniques described in earlier chapters to solve a distributed disturbance in an industrial production plant at the Eastman Chemical Company, Kingsport Tennessee, USA. It gives an opportunity to demonstrate the portability of the methods to a process different from the one in which they were originally tested as well as to solve an real industrial problem for the Eastman Chemicals collaborators. Topics of interest include:

- Visual examination of time trends, autocovariance functions and spectra;
- Spectral principal component analysis to detect disturbances having similar spectra;
- Detection of multiple oscillations;
- Non-linearity analysis to determine time trends originating from non-linear elements in the plant;
- Suggestions of the origins of disturbances.

8.1.2 Introduction

The chapter starts with a description of the process and a survey of the process data provided for the study. It then describes the detection of plant-wide disturbances and groups of oscillations. The oscillation detection tools and spectral PCA were used for this purpose.

Section 8.3 presents the diagnosis of the root causes of the disturbances. The most serious disturbance was caused by a sticking level control valve which was successfully identified from its non-linearity index. The finding was a new diagnosis; the problem with this control valve was not known in advance. The Eastman control engineers conducted some invasive tests on the valve and confirmed the finding from the non-linearity assessment. The root causes of other detected disturbances are also discussed.

The analysis was successful and resulted in maintenance action at the next shutdown that cured a major plant-wide disturbance. Data from before and after the shutdown are presented to demonstrate the improvement.

8.1.3 Work done by the thesis author and others

The work reported in this chapter was first presented as an internal report to Eastman Chemical Company in 2001. The following paper was published on the case study:

> Thornhill, N.F., Cox, J.W., and Paulonis, M., 2003, Diagnosis of plant-wide oscillation through data-driven analysis and process understanding, *Control Engineering Practice*, 11, 1481-1490.

The report and paper were written by Thornhill who also created the software, results, graphs and all but one of the diagrams. The diagram below showing the process schematic was drawn and provided by John Cox. The co-authors also selected the case study, provided data, process insights and understanding, and performed additional valve and controller tests at Kingsport.

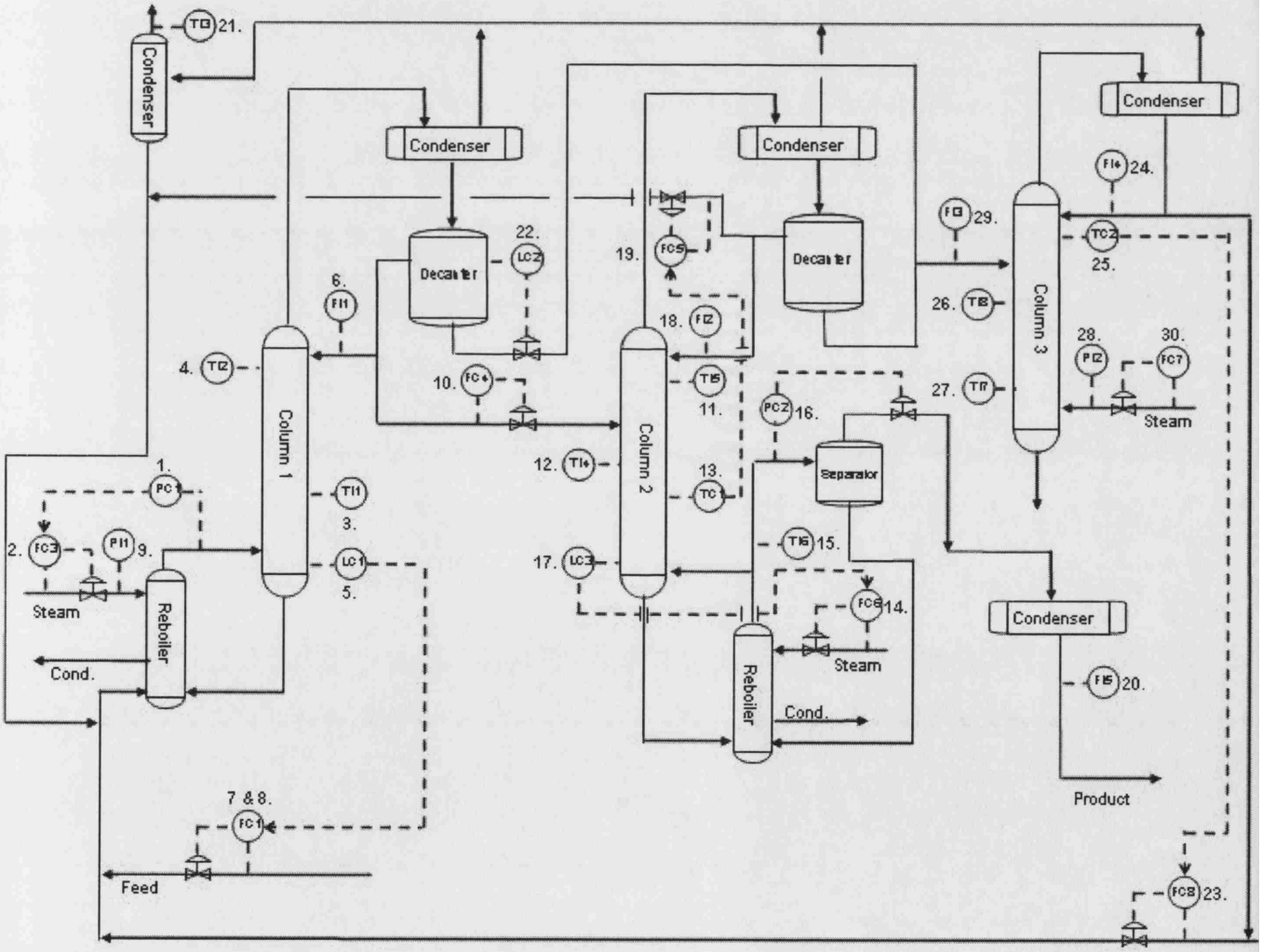

Fig 8.1 Process schematic for the industrial case study (diagram courtesy of J.W. Cox, Eastman Chemical Company).

8.2 DETECTION OF PLANT-WIDE DISTURBANCES

8.2.1 Review of the process and its data

The process schematic: The process schematic and data were provided by John Cox of Eastman on May 8th 2001. The data were 20s samples from two days running comprising 15 control loops and 15 instruments from a three column separation unit. Two decanters and a separator were also in use. The process is shown in Figure 8.1 where the following structural features can be observed:

- Feed to columns 2 and 3 from the top stream of column 1;
- Recycles of material from columns 2 and 3 to column 1;
- Direct steam injection heating to column 3

The data suggested the presence of multiple plant wide disturbances. The purpose of this report is to show an analysis of the disturbances.

Basic statistics and time trends: The means and standard deviations were determined for all the measured process variables (*pv*). These quantities are in engineering units and are company confidential therefore they cannot be presented here. The effect of plant-wide disturbances is that the product flow rate leaving the plant (FI5, Tag 20) deviated by ±4.3% of its mean value.

Presenting the values as a percentage of set point can hide the engineering units, as was done with FI5 above. However, this practice can be very misleading in cases where the scale has a false zero such as is the case with temperature or pressure measurements where the zero of the instrument range is not $0\,\mathrm{K}$ or $0\,\mathrm{Pa}$. The ideal way to present measurements without revealing the true values is to express then as a percentage of the range of the instrument. The instrument ranges were not available for the Eastman data sets, however.

Normalized time trends of 30 process variables or measured variables (*pv*), the set points (*sp*) and controller outputs (*op*) for the 15 controllers in the data set in Figures 8.2 to 8.4. Normalization means that each was detrended and scaled to unit standard deviation before plotting.

Compression analysis: A compression analysis was conducted on the *pv*'s of the Eastman data. None of them was significantly compressed and all are analyzable because all the compression factors were below the $CF = 3$ threshold.

Tag	CF	Tag	CF	Tag	CF
1	1.0	11	1.5	21	1.9
2	1.0	12	1.2	22	1.0
3	2.4	13	1.1	23	1.2
4	1.5	14	1.0	24	1.0
5	1.0	15	1.4	25	1.6
6	1.0	16	1.0	26	1.6
7	1.0	17	1.0	27	2.1
8	1.0	18	1.0	28	1.0
9	1.1	19	1.0	29	1.0
10	1.0	20	1.0	30	1.0

Table 8.1 Compression factors of the Eastman data.

8.2.2 Spectral principal component analysis

Spectra and autocovariance functions: Figure 8.5 shows the power spectra of the process variables (*pv*). The *pv* have been mean centred before the DFT operation was applied to remove d.c. or zero frequency components. The spectra have also been scaled for the purposes of visualization so that the maximum peak height is the same in each. The reason for this scaling is that it makes it easier to see broad-band spectra such as those of Tags 1 and 10. When the spectra are used in a spectral PCA calculation, however, they are re-normalized to unit power. Figure 8.6 shows the autocovariance functions (ACF) which are normalized so that the ACF at zero lag is equal to 1.

Several spectral peaks and oscillating ACFs are evident, the most prominent spectral peak is just under 0.03 on the normalized frequency axis corresponding to a oscillation with a period of about 340 samples per cycle in the time domain and ACF. Since the sampling interval was 20s, this is the oscillation with a period of just under two hours. It can be seen easily in some of the time trends such as Tags 12, 13, 19 and 22. Other smaller peaks are also present in the spectra with rapid oscillations in the time trends and ACF and some tags such as 16, 17 and 20 are subject to multiple oscillations.

The challenges for analysis are:

- To detect and characterize all plant-wide disturbances;
- To find which tags are affected by the disturbances;
- To determine the root cause(s) of the disturbances.

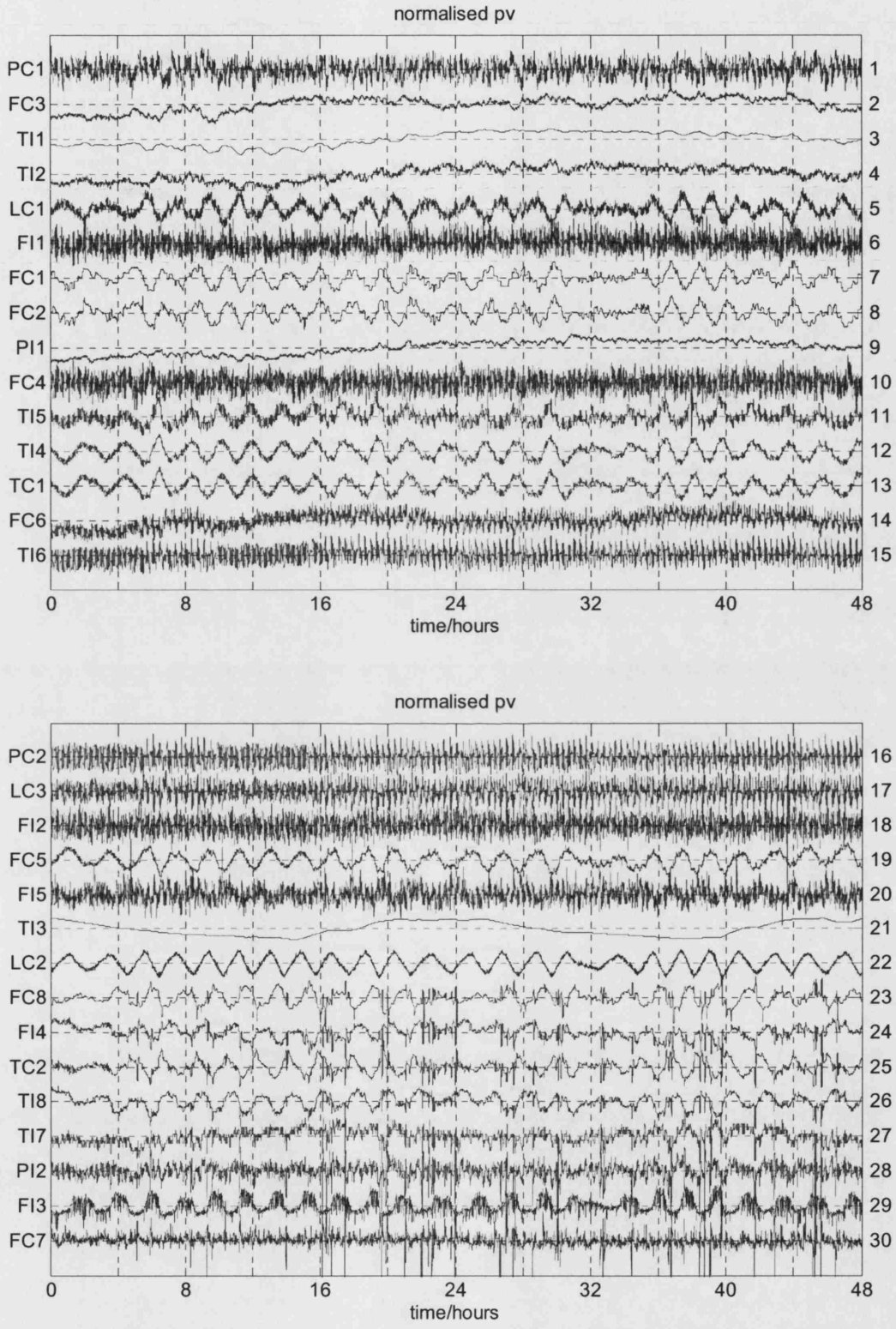

Fig 8.2 Process variables (*pv*) for the industrial case study. The trends are mean centred and scaled to unit standard deviation.

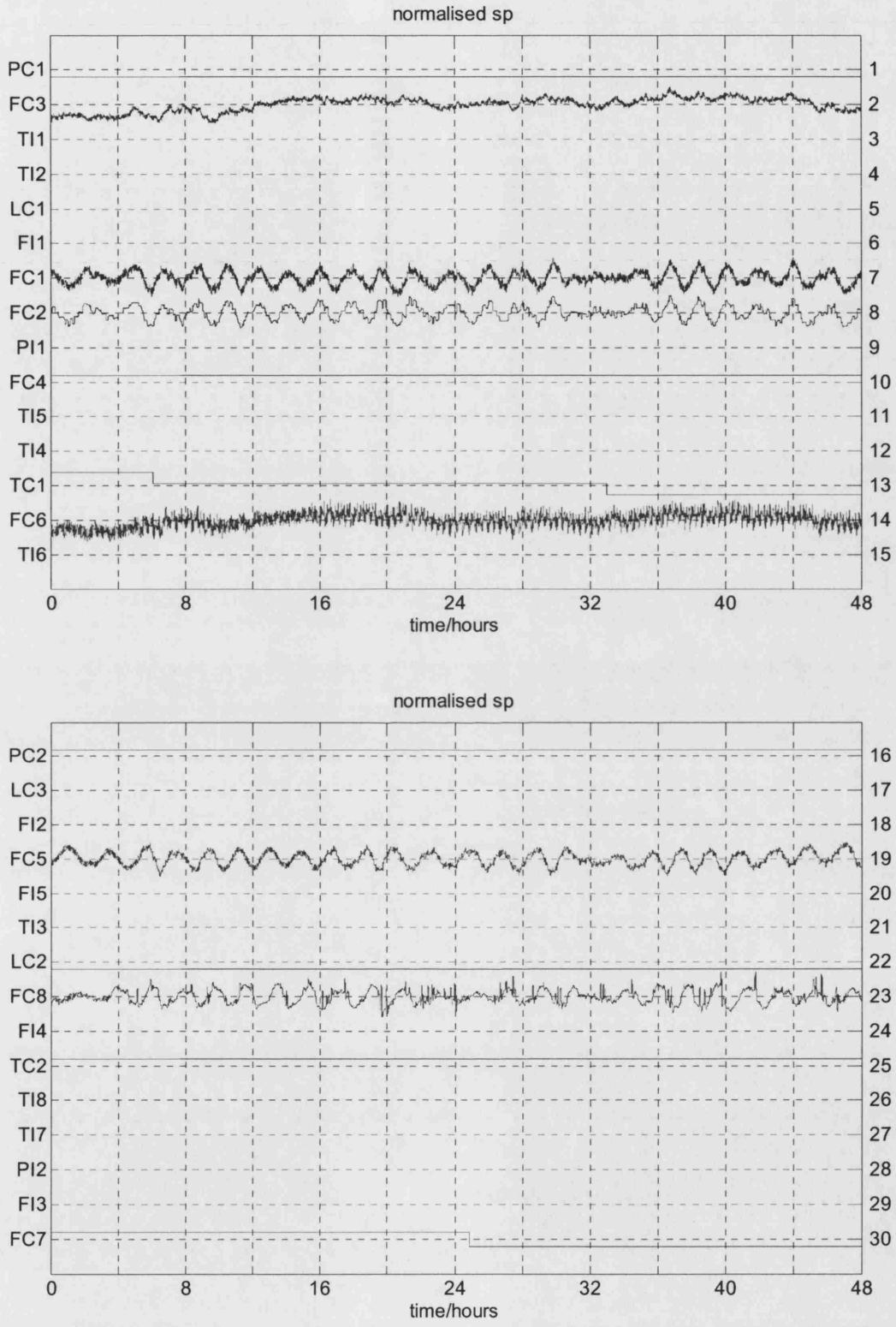

Fig 8.3 Set points (*sp*) for the 15 controllers in the industrial case study. The trends are mean centred and scaled to unit standard deviation.

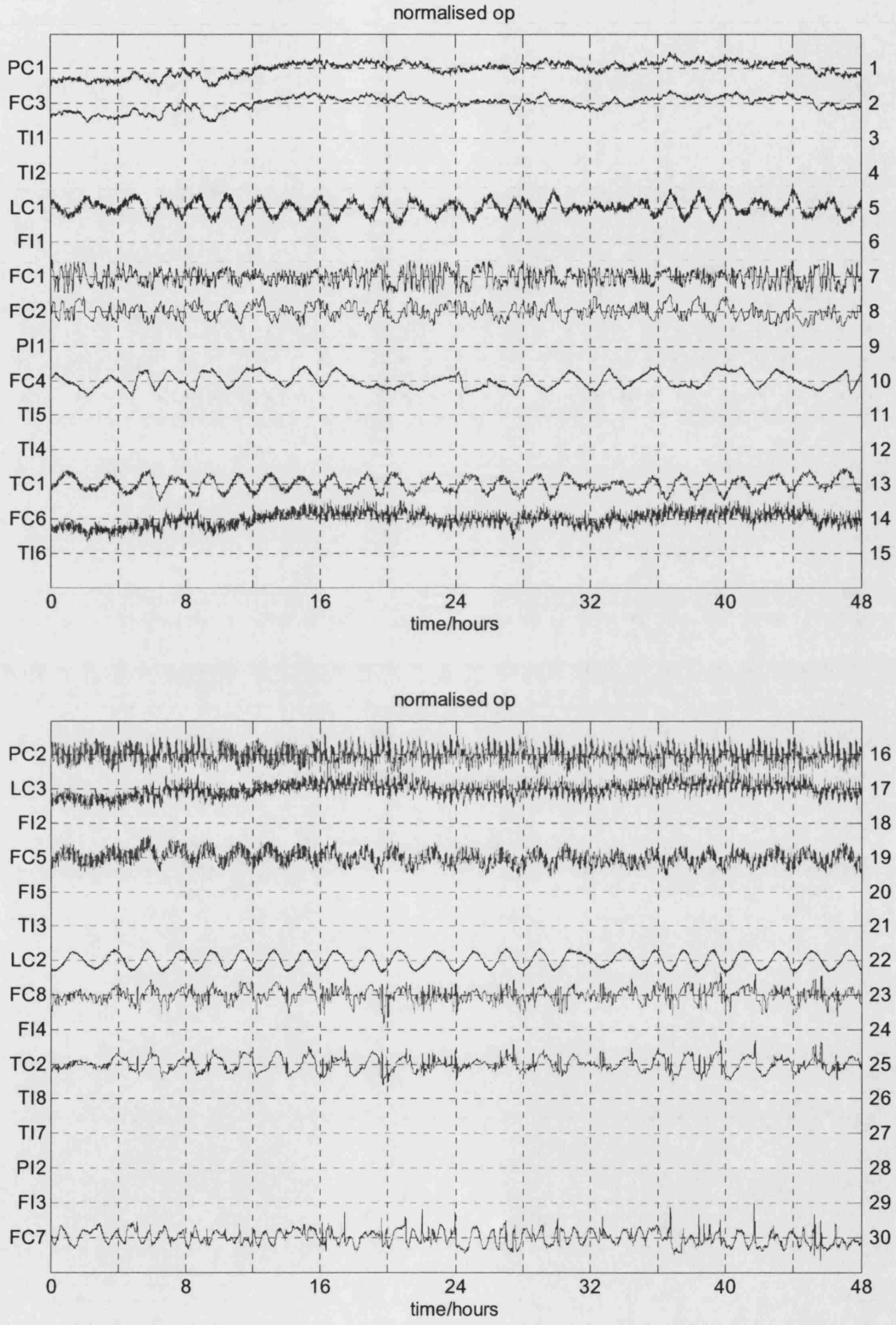

Fig 8.4 Controller outputs (*op*) for the 15 controllers in the industrial case study. The trends are mean centred and scaled to unit standard deviation.

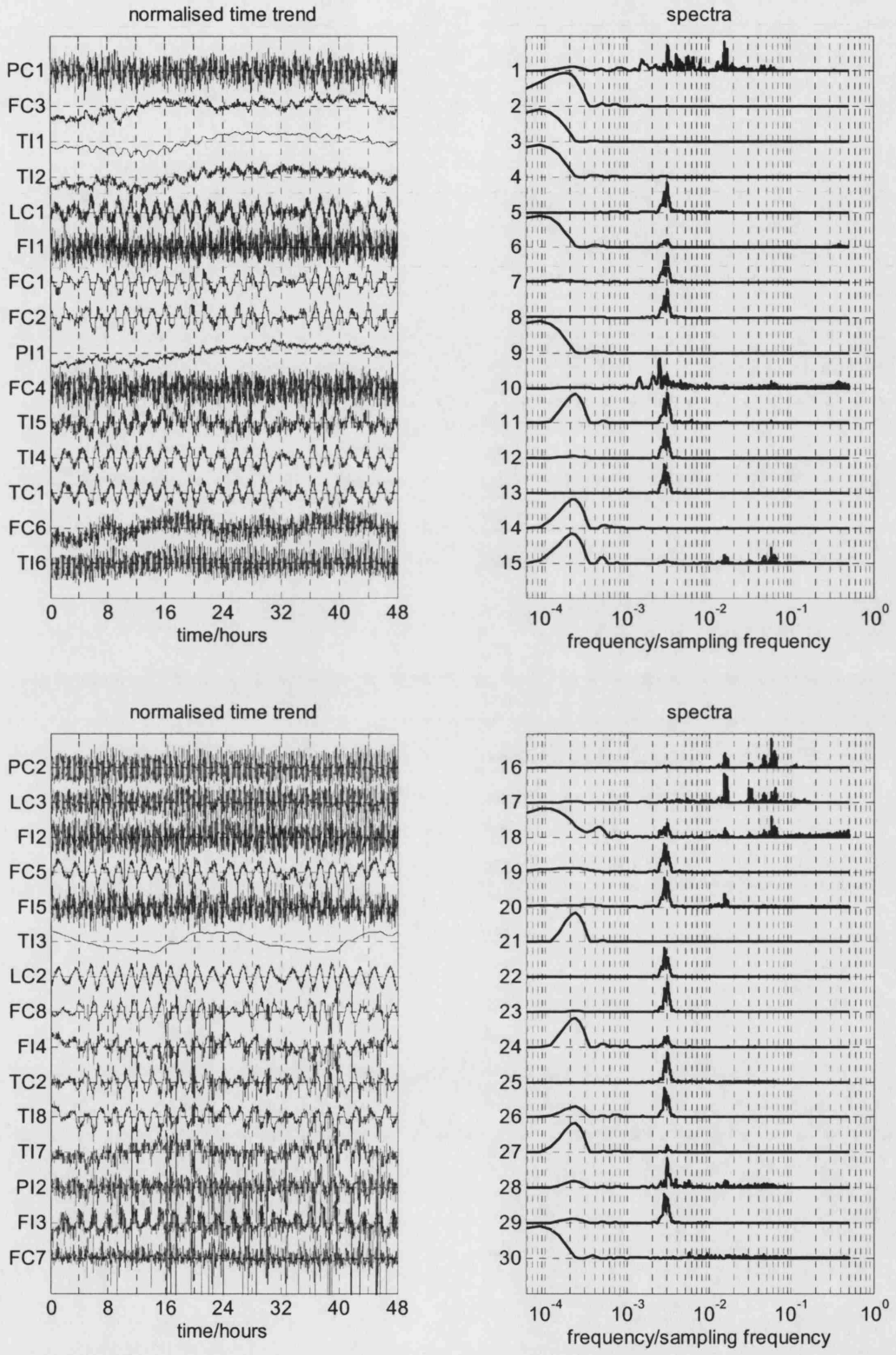

Fig 8.5 Process variables (*pv*) and spectra. The *pv*'s are mean centred and scaled to unit standard deviation. The spectra have been scaled to the same maximum peak height.

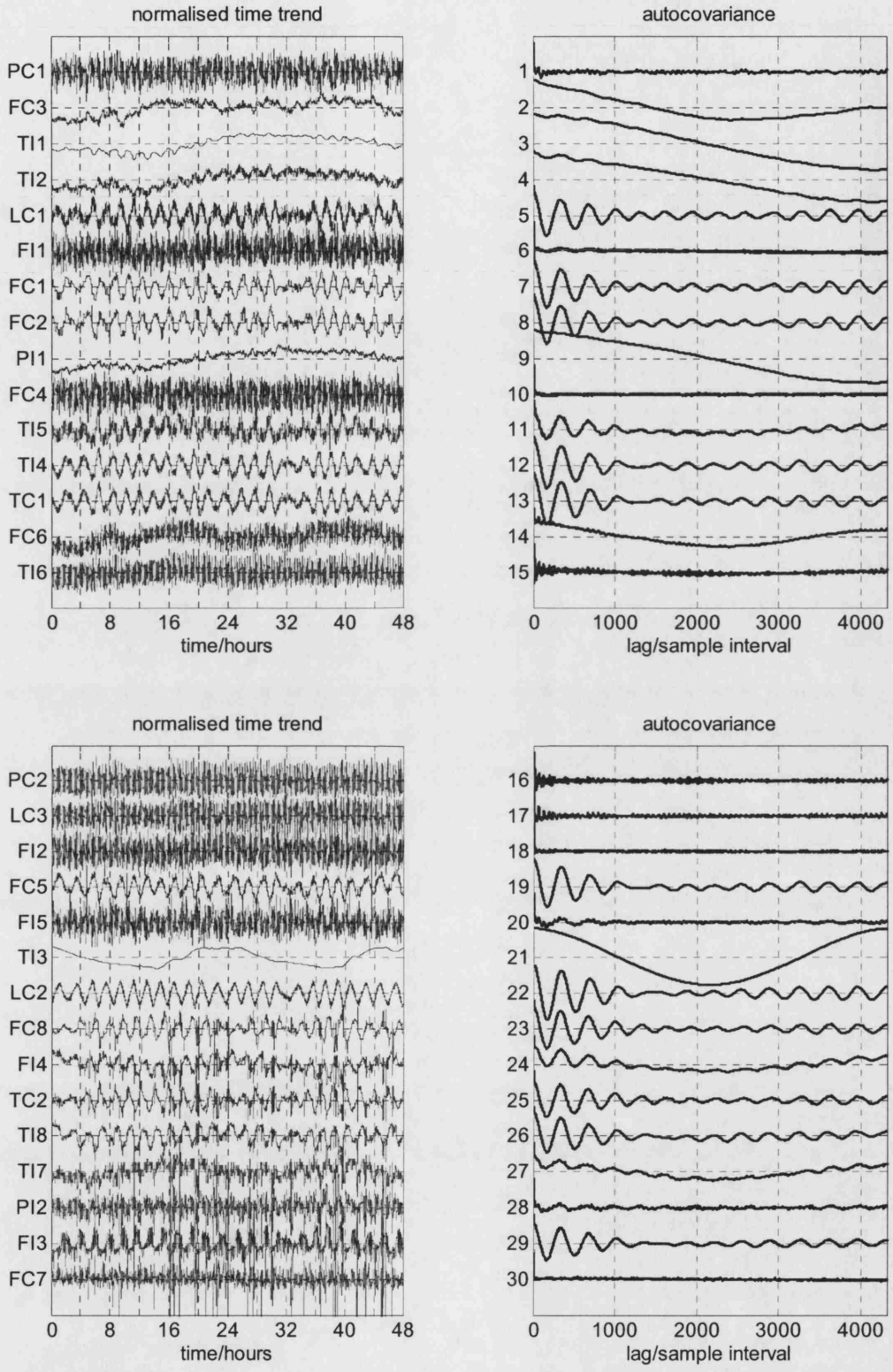

Fig 8.6 Process variables (*pv*) and autocovariance functions (*ACF*) The *pv*'s are mean centred and scaled to unit standard deviation. The *ACF* are normalized to unity at zero lag.

Spectral PCA: A spectral principal component analysis was carried out in order to detect and characterize the main disturbances present in the plant. Eight principal components captured 99.75% of the variability in the data set and therefore 8 PCs were used to construct the hierarchical classification tree shown in Figure 8.7.

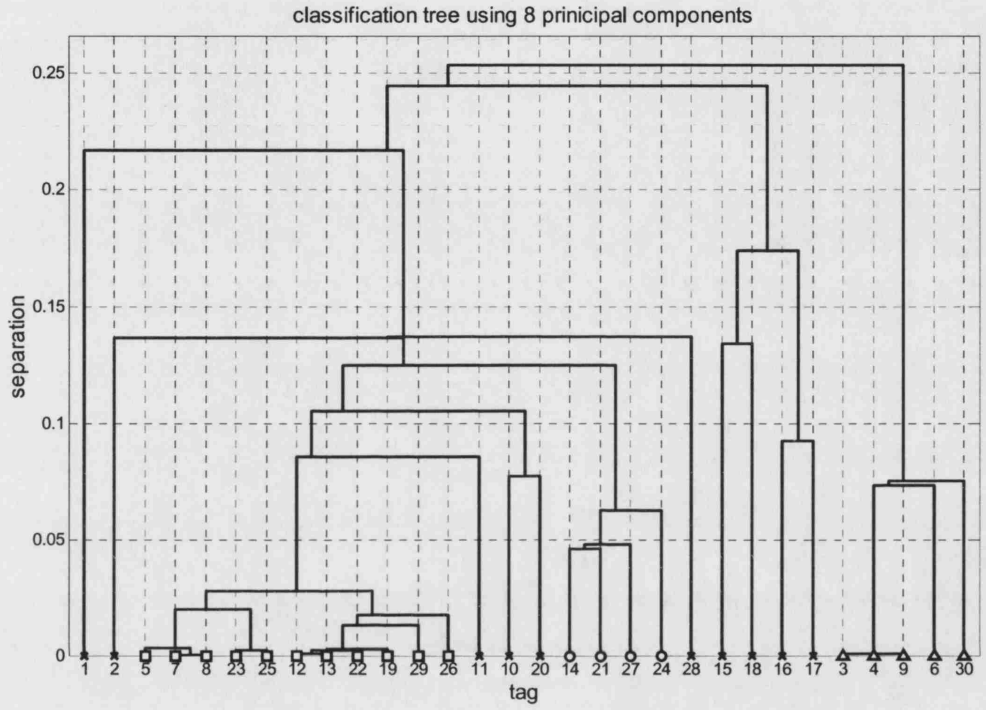

Fig 8.7 Spectral hierarchical classification tree for the Eastman Chemicals process.

Disturbances detected by spectral PCA: Figure 8.7 shows the main groups of disturbances identified by spectral PCA. A groups of tags are considered to form a cluster when they are all joined to each other by short twigs that are all shorter then the main branch that joins the group to the rest of the tree. By reference to the time trend, spectral and autocovariance plots the disturbances can be characterized as follows:

- Tags with O symbols (14, 21, 24 and 27). These have diurnal (24 hour) variations. The fact that these are 24 hours may be determined from the spectra. The peak on the normalized frequency axis is at 0.0023. Therefore there are 1/0.00023 or 4348 20s samples per cycle, i.e. $4348/(3\times 60) = 24.15$ hours.

  These variations are caused by daily temperature cycles and are mainly affecting column 3 (Tags 21, 24 and 27; TI3, FI4 and TI7). It is possible that Column 3 is near the top of the building housing the plant and therefore more directly affected by ambient temperature changes than units lower down in the cooler part of the building. TI3, FI4 and TI7 are all indicators and are not controlled variables.

Tag 14 (FC6) is the steam flow of the reboiler of Column 2. It has a daily cycle in temperature because the set point of the FC6 control loop is demanding such behaviour. Because of the level-flow cascade structure, FC6.SP is the same as LC3.OP (Tag 17) and therefore the reason why FC6 varies is to compensate for the effects of daily cyclic disturbance on the level at the bottom of column 2.

Daily temperature cycles are observed in most processes and are not normally of great significance in the operation of the process. Therefore the disturbance characterized by circles will not be discussed further.

- Tags indicated by △ symbols (3, 4, 6, 9 and 30). The spectra show that these cycles last longer than 24 hours because the spectral peak is even further towards the low frequency end of the spectra than those denoted by ○. The time trends and ACF plots show that Tags 3, 4, 6 and 9 all share a common slow deviation. Tag 30 does not share the same behaviour at all and its grouping with 3, 4, 6 and 9 is a spurious artifact of the fact that their spectra have very low frequency components. The reason why Tag 30 has a very low frequency feature is because of a step-change in its set point just after 24 hours (Figure 8.3)

  The very slow disturbance is associated with column 1 (Tags 3, 4, 6 and 9; TI1, TI2, FI1 and PI1). All are indicators and not under closed loop control. There is no indication of the disturbance in the feed flow rate (Tags 7 and 8) nor in the steam flow rate. Since the thermodynamic properties of temperature and pressure are involved it is possible that the disturbance is due to gradual changes in composition of the feed.

- Tags indicated by □ symbols (5, 7, 8, 12, 13, 19, 22, 23, 25, 26, and 29) have a persistent oscillation that appears at about 0.003 on the normalized frequency axis. Therefore the red oscillations have 330-340 samples per cycle and a period of just under two hours. These oscillations show also in the autocovariance functions.

  The oscillation is very widespread and it is not at all clear from visual inspection where it originates. It is not possible to suggest a root-cause at this stage and further investigation will be needed.

- Tags 15, 16, 17 and 18 are shown to have spectra closer to each other than to any other spectra but they do not form a cluster. Their similarity is noted, however. They have a complicated set of high frequency features. The spectra suggest that they have oscillatory features at about 0.015 and 0.055 on the normalized frequency axis, i.e. about 66 and 18 samples per cycle.

The presence of multiple oscillation will require a further step in the analysis (see section 8.3) because the oscillations may also be present at a low level in some of the tags but masked by the other disturbances described above. The distribution of the 66 and 18 samples per cycle oscillations will become more clear when frequency domain filtering has been applied to remove the interfering effects.

Summary: Four distributed disturbances have been detected and characterized by spectral PCA. The most widespread is an oscillation with a period of about 333 samples per cycle which is present in many parts of the plant.

Other oscillatory disturbances with periods of about 66 and 18 samples per cycle are also present but these may be masked by other disturbances in some tags. Therefore frequency domain filtering will be needed to determine the extent of their distribution in the plant.

Diurnal variations have been observed in some indicators mainly associated with Column 3 and very slow deviations localized in column 1 have been attributed to an external disturbance, possibly changes in composition of the feed.

8.2.3 Oscillation detection

Introduction: The spectra showed multiple spectral peaks and therefore the an automated analysis for detection of multiple oscillations was applied as described in Chapter 4. A threshold for detection of 10% power was applied so that insignificant oscillations would not be reported.

Application of automated oscillation detection:

Results

Step 1: The filter used at step 1 was 2 to 1024 samples per cycle. Any oscillations up to 1024 samples per cycle would have at least eight zero crossings in the unaliased part (i.e. lags 0 to 4196) of the autocovariance function. The lower bound on the filter was 2 samples per cycle (the Nyquist criterion).

Step 2: The method for counting and grouping the detected oscillations reported the following oscillations having regularity factors $r > 1$ and more then 10% power in the selected frequency range:

- Two tags, average $\bar{T}_p$ of 17.6 samples per cycle;
- Ten tags, average $\bar{T}_p$ of 349 samples per cycle;

Step 3: Two oscillations were detected so the algorithm generated two new filters to refine the search:

- 4 to 33 samples per cycle;
- 33 to 521 samples per cycle.

The following oscillations were detected:

- Four tags, average $\bar{T}_p$ of 17.6 in filter 1;
- Five tags, average $\bar{T}_p$ of 63 in filter 2;
- Thirteen tags, average $\bar{T}_p$ of 344 in filter 2.

Step 4: The above result for filter 1 was final because only one oscillation was detected in its range. Step 2 was repeated for filter 2 because there were two oscillations in its range. The two new filters were:

- 43 to 106 samples per cycle;
- 106 to 414 samples per cycle.

The final result for the sub-divided filter 2 was:

- Seven tags, average $\bar{T}_p$ of 64;
- Sixteen tags, average $\bar{T}_p$ of 341.

The findings are summarized in Table 8.1.

slow: 106 to 414 sample/cycle					medium: 43 to 106 sample/cycle				fast: 4 to 33 sample/cycle			
name	tag	period	*r*	power	tag	period	*r*	power	tag	period	*r*	power
PC1	1				1	63±2.6	8.0	23%	1			
FC3	2				2				2			
TI1	3				3				3			
TI2	4				4				4			
LC1	5	320±24	4.6	80%	5				5			
FI1	6				6				6			
FC1	7	322±30	3.6	83%	7				7			
FC2	8	321±31	3.5	83%	8				8			
PI1	9				9				9			
FC4	10				10				10			
TI5	11	316±41	2.6	55%	11				11	17.3±2.6	2.2	23%
TI4	12	381±82	1.5	87%	12				12			
TC1	13	383±92	1.4	87%	13				13			
FC6	14				14				14			
TI6	15				15	64±1.6	14.0	13%	15	17.6±3.1	1.9	75%
PC2	16				16	64±1.4	15.0	11%	16	17.6±3.1	1.9	82%
LC3	17				17	64±1.3	16.0	18%	17			
FI2	18				18				18	18.0±3.2	1.9	41%
FC5	19	382±78	1.6	83%	19				19			
FI5	20	314±57	1.8	15%	20	64±2.5	8.4	16%	20			
TI3	21				21				21			
LC2	22	380±81	1.6	95%	22				22			
FC8	23	329±15	7.4	77%	23				23			
FI4	24	330±30	3.7	39%	24				24			
TC2	25	329±16	7.0	69%	25				25			
TI8	26	377±87	1.4	71%	26				26			
TI7	27	327±20	5.6	26%	27	63±4.6	4.6	11%	27			
PI2	28	329±27	4.1	21%	28	64±4.4	4.8	16%	28			
FI3	29	318±64	1.7	65%	29				29			
FC7	30				30				30			

Table 8.2 Oscillations detected in the Eastman Chemicals data set.

Oscillation characterization: Table 8.2 shows there were three plant-wide oscillations:

- Slow oscillation: Tags 5, 7, 8, 11-13, 19, 20, 22-29, average period 341 samples (114 min);
- Medium oscillation: Tags 1, 15-17, 20, 27, 28, average period 64 samples (21.3 min);
- Fast oscillation: Tags 11, 15, 16, 18, average period 17.6 samples per cycle (5.87 min).

Tags 20, 27, 28 had slow and medium oscillations present, Tags 15 and 16 have medium and fast oscillations while Tag 11 has slow and fast oscillations. Therefore the procedure successfully identified the presence of multiple oscillations, both plant-wide cases where the same oscillation was present in more than one tag and cases where some tags had more than one oscillation.

Distributions of the disturbances: The distribution of the slow oscillation at 341 samples per cycle is shown in Figure 8.8. The disturbance was a serious one because the signal power associated with that oscillation was high, more than 80% in seven of the tags and 90% or more in one of them. The oscillation is very widespread and it is not possible without more analysis to determine in which unit it originates.

The medium plant-wide oscillation at 64 samples per cycle also affected many tags at different places in the plant (Figure 8.9) but the signal power was smaller, the highest being 23% in tag 1. The location of this disturbance near the bottom of all three columns strongly suggests a disturbance entering from the steam system.

The fast oscillation at 17.6 samples per cycle spread to four tags and the associated power was high in two of them (15 and 16). All are associated with Column 2, as shown in Figure 8.10 and this disturbance is therefore quite localized.

The majority of the variability in the plant was due to the slow oscillation at 341 samples per cycle . Its power was high and many tags are involved, and priority should be given to its analysis.

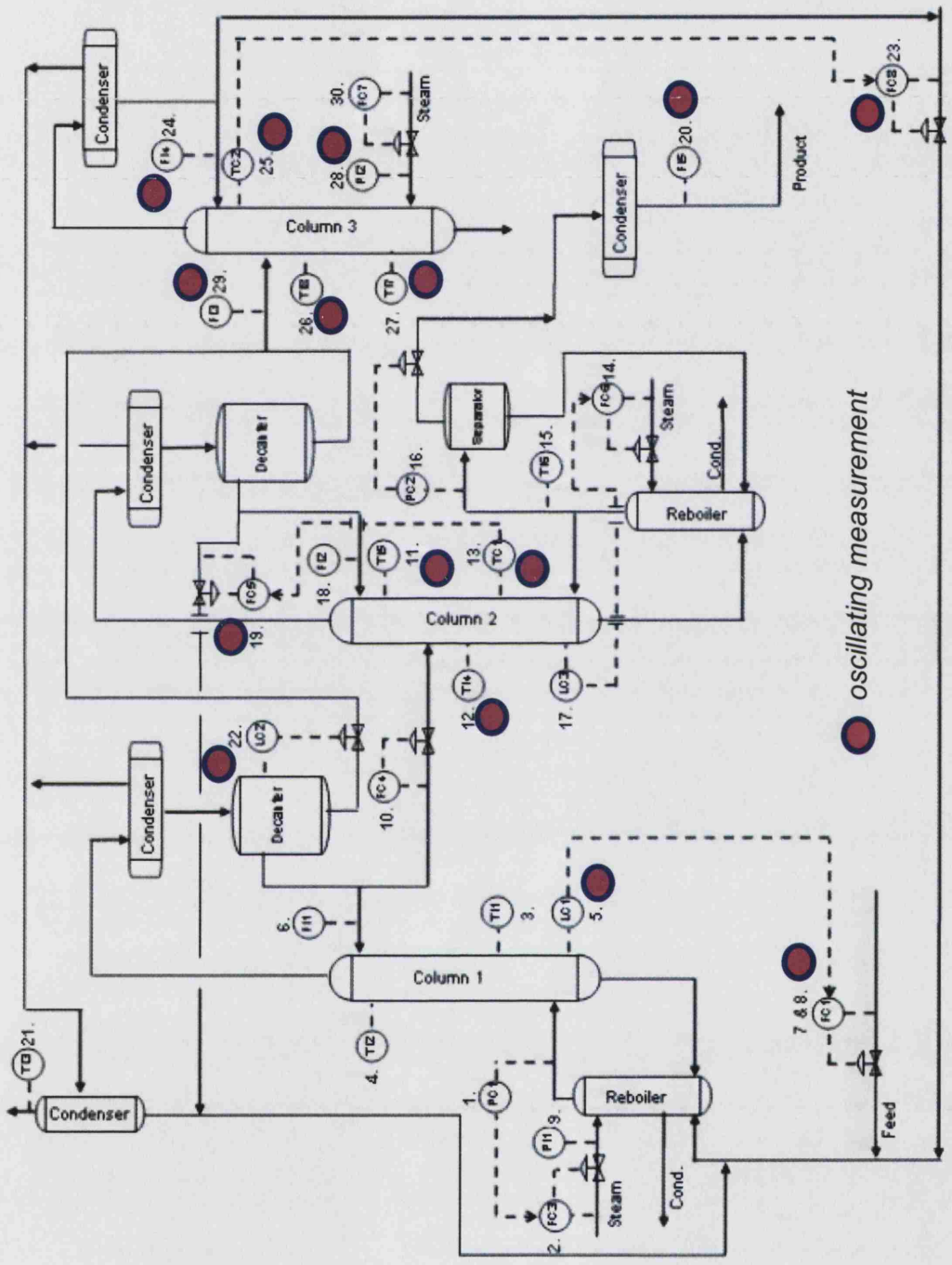

Fig 8.8 Distribution of the slow oscillation with an average period of 341 samples per cycle.

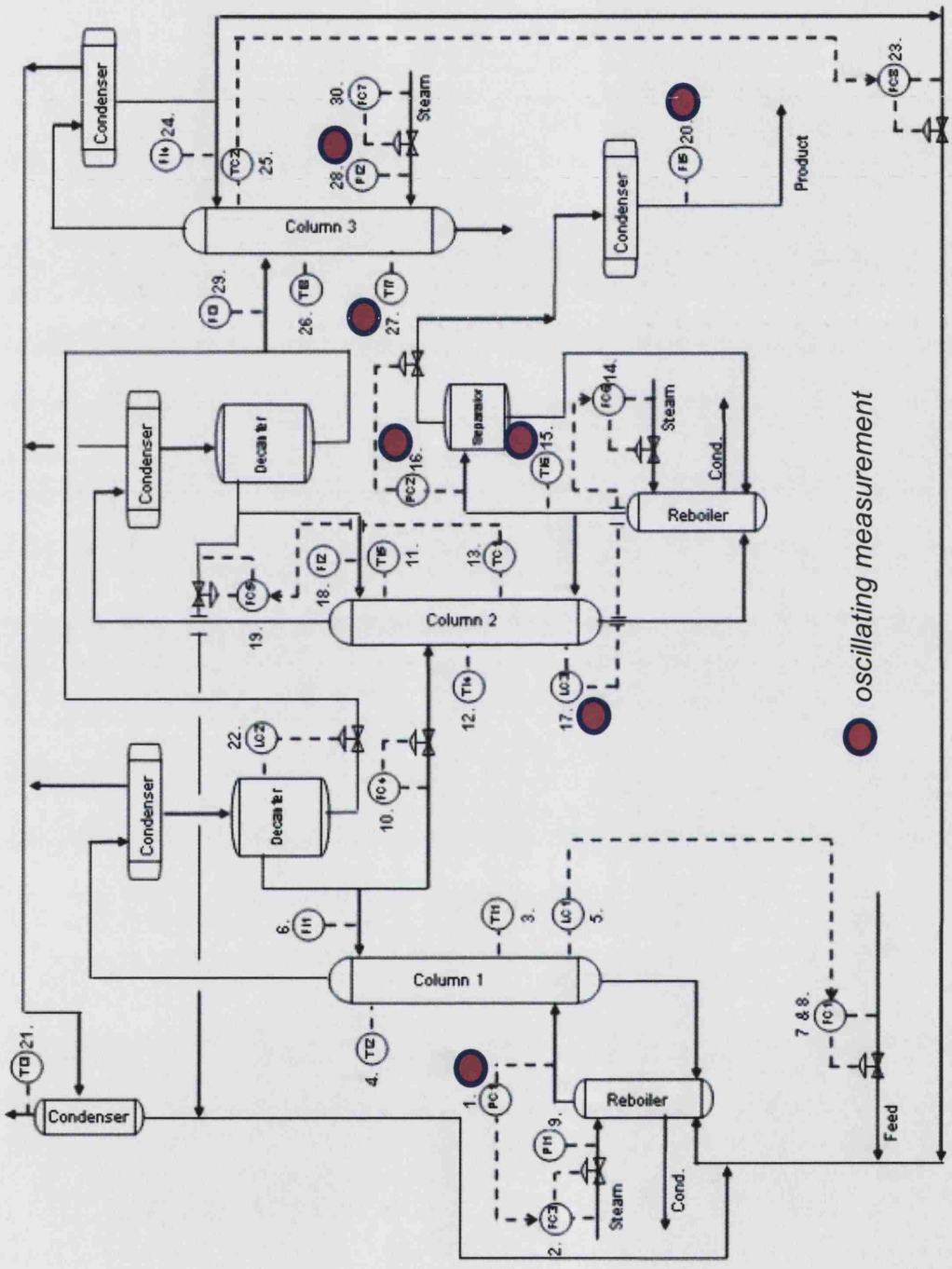

Fig 8.9 Distribution of the medium oscillation with an average period of 64 samples per cycle.

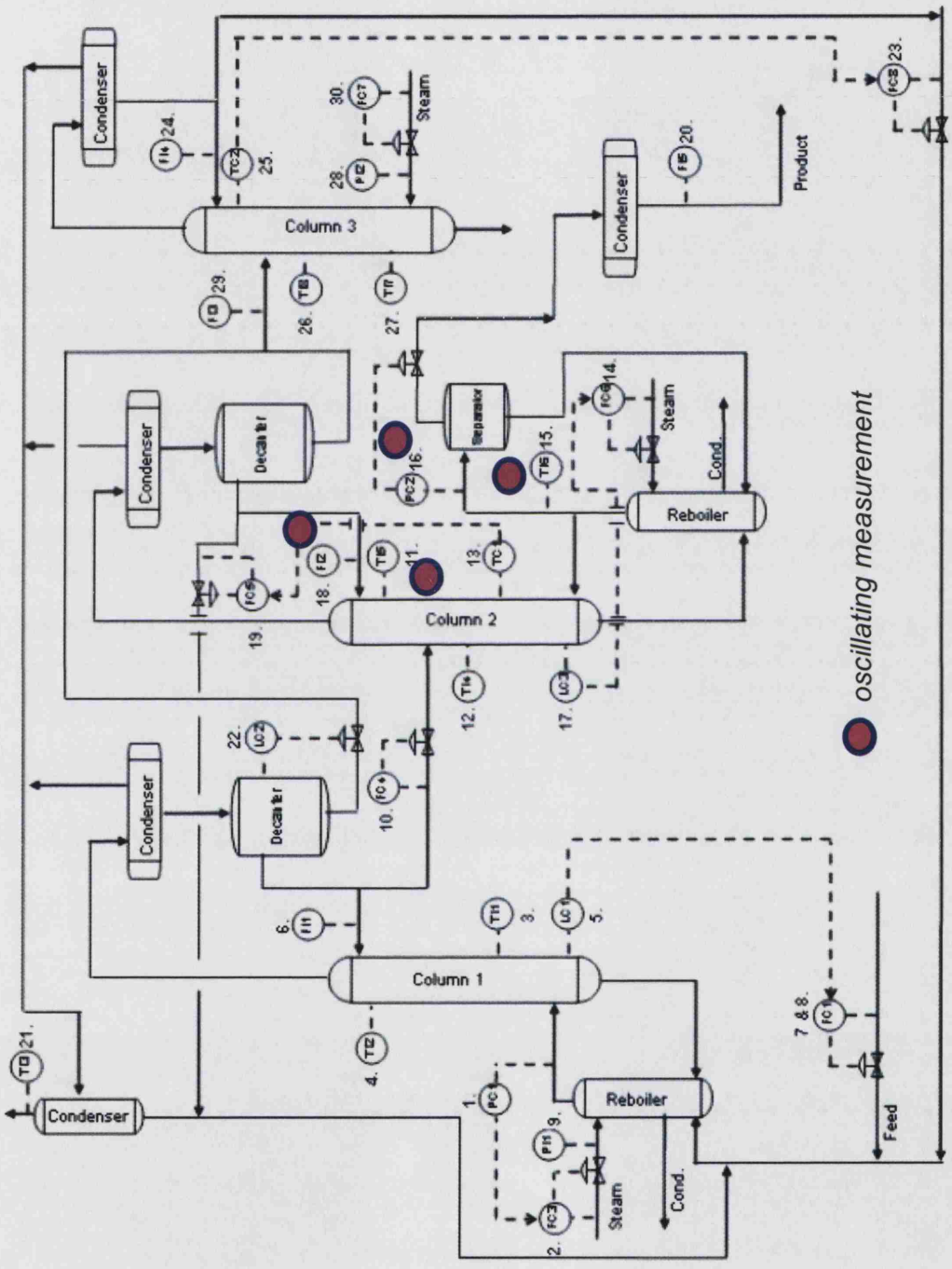

Fig 8.10 Distribution of the fast oscillation with an average period of 17.6 samples per cycle.

8.3 DIAGNOSIS OF PLANT-WIDE DISTURBANCES

8.3.1 Diagnosis of the medium and fast plant-wide oscillation

Timing of the oscillations: The 17.6 and 64 samples per cycle oscillations are not present all the time in the data set. The 17.6 samples per cycle oscillation comes and goes sporadically during the first 20 hours and alternates with the 64 samples per cycle oscillation. Figure 8.11 shows that tags 15 and 16 which show the strongest participation in the 17.6 samples per cycle oscillation alternate between that behaviour and the 64 samples per cycle oscillation. At the beginning the fast 17.6 samples per cycle oscillation is present, Then from 5.5 hours to 10 hours the 64 samples per cycle oscillation takes over followed by another episode of the 17.6 samples per cycle oscillation. After 16 hours the 64 samples per cycle oscillation returns and stays in place until the end of the 48 hour data set.

The conclusion is that the oscillations are related to one another because they appear consecutively in the same tags. Normally in such a case one seeks for another variable such as plant throughput that might explain different dynamics at different operating points but in this case there are no obvious movements in set point that correspond to the changes in the period of oscillation. The change has therefore been noted but an explanation could not be discovered within the data set.

Non-linearity analysis: Samples 100 to 400 (about 0.5 to 2 hours) were used for the diagnosis of the 17.6 samples per cycle oscillation (period of about 0.1 hour) in tags 11, 15, 16 and 18. A pre-filter was applied to remove any frequency components with periods longer than 100 samples per cycle so that, for instance, the 341 sample oscillation would be removed from tag 11. Other conditions were: $E = H = 17$, $k = 8$ and 50 surrogates. The results are shown in the upper panel of Figure 8.12.

Samples 5000 to 5900 (27.7 to 32.7 hours) were used for analysis of the 64 samples per cycle (21 minutes or 0.35 hour) oscillation in tags 1, 15, 16, 17, 20, 27 and 28. The data were sub-sampled by 3 to give roughly 20 to 21 samples per cycle for the non-linearity analysis. Other conditions were $E = H = 20$ $k = 8$ and 50 surrogates. The results are shown in the lower panel of Figure 8.13.

The shapes of the plots in Figure 8.13 qualitatively support the findings on non-linearity. The waveforms of tags 15 and 16 in the upper panel are distinctly sawtooth in shape suggesting a non-linear origin. In the lower panel the 64 samples per cycle has a very prominent 3rd harmonic, in fact it is sufficiently prominent in some of the waveforms as to mask the fundamental. A non-sinusoidal waveform is again indicative of generation by a non-linear process.

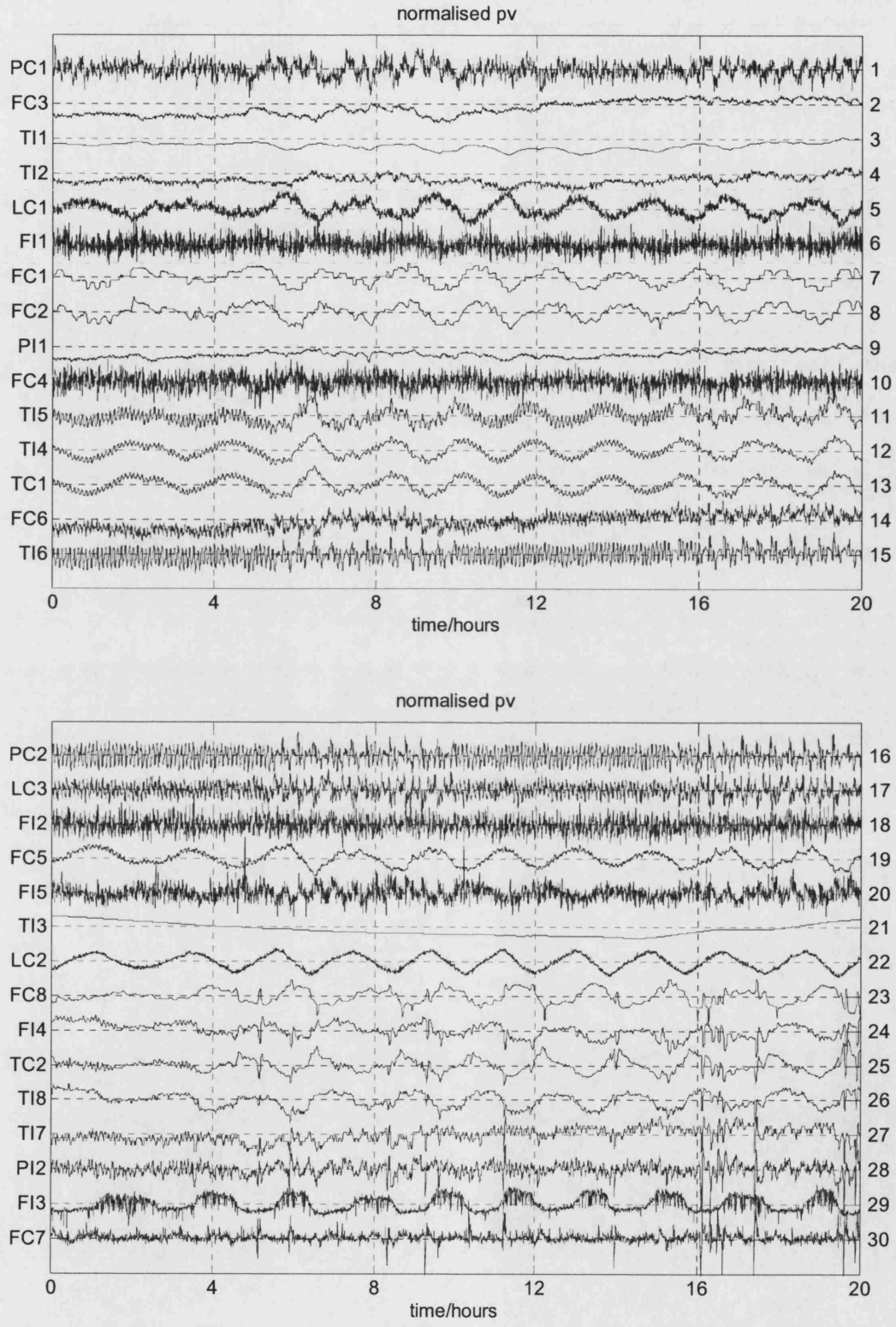

Fig 8.11 Close up of the first 20 hours of the Eastman data set showing that Tags 15 and 16 alternate between oscillation episodes having fast and medium periods.

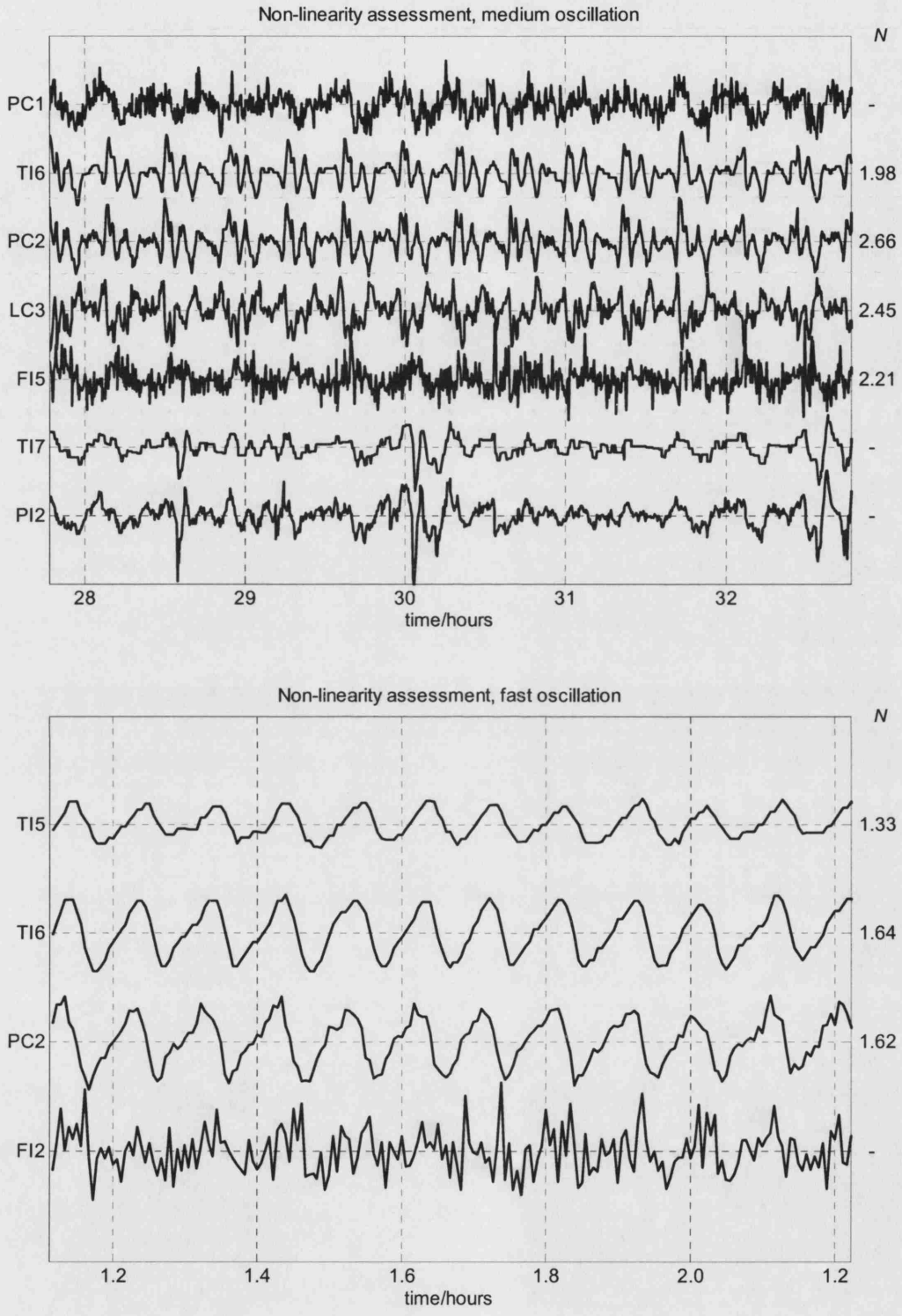

Fig 8.12 Non-linearity assessments. Upper panel: The 64 samples per cycle oscillation (0.35 hour). This oscillation has a prominent third harmonic in PC2 (Tag 16) and TI6 (Tag 15). Lower panel: The 17.6 sample per cycle oscillation period (0.1 hour).

Oscillation/samples	samples selected	filter/samples per cycle	subsampling	E, H
17.6 (section 8.3.1)	100 – 400	2 – 100	none	17
64 (section 8.3.1)	5000 – 5900	2 – 100	3	20
341 (section 8.3.2)	720 – 5000	100 – 1024	10	30

Table 8.3. Settings for non-linearity analysis of the fast, medium and slow oscillations.

17.6 sample oscillation (fast)			64 sample oscillation (medium)		
Name	Tag No	*N*	Name	Tag No	*N*
TI5	11	1.33	PC1	1	–
TI6	15	1.64	TI6	15	1.98
PC2	16	1.62	PC2	16	2.66
FI2	18	–	LC3	17	2.45
			FI5	20	2.21
			TI7	27	–
			PI2	28	–

Table 8.4 Non-linearity analysis of the fast and medium oscillations.

Root cause of the medium oscillation: The medium oscillation with 64 samples per cycle is shown in the upper panel of Figure 8.12. Two instruments and two controlled variables in the Column 2 reboiler area were non-linear. PC2 (Tag 16) had the highest non-linearity at 2.66 and the non-linearity of LC3 (Tag 17) was also high. Therefore these control loops are investigated as possible root causes.

Both controller output (LC3.OP, Tag 17) and the flow through the control valve (FC6.PV, Tag 14) are available measurements and are plotted in Figure 8.13 on an *op-mv* plot (upper panel) and as time trends (lower panel). The 45° angle of the *op-mv* plot shows the FC6 steam flow control valve is healthy and is responding in a linear manner to the LC3 controller output. The FC3.PV flow measurement is doing exactly what the LC3.OP (Tag 17 *op*) and FC6.SP (Tag 14 *sp*) are asking it to do. LC3 (Tag 17) is not the root cause, therefore.

There is no measurement of flow through the PC2 (Tag 16) control valve and therefore it is not possible to create an *op-mv* plot for this control loop. Figure 8.14 plots instead the *op-pv* plot for Tag 16. It is noted in Figure 8.14 that the PC2.OP and PC2.PV appear to move together. This observation is ambiguous, however, and cannot be used to conclude that the PC2 control valve is healthy. For instance, those two signals would be identical if the PC2 controller is proportional-only regardless of the state of the valve so the similarity merely indicates that the controller has light integral action.

In conclusion, the source of non-linearity in the 64 samples per cycle oscillation can be localized to the bottom end of Column 2. It is not caused by the FC6 (Tag 17) steam flow valve which is

behaving linearly, but non-linearity of some type is certainly present. The high non-linearity in PC2 (Tag 16) indicates that it would be beneficial to investigate further.

Certainly the PC2 control valve should be tested. However, the problem might be a process non-linearity rather than a valve fault. It is possible that mild oscillations in the steam pressure are triggering a non-linear response PC2 pressure control loop (Tag 16) in the Column 2 reboiler area. For instance, if the hold up in the reboiler is small, and if the container of the liquid has sloping sides then the liquid-vapour interface area would change with level. The reason for thinking the steam system may be bringing an oscillation into the plant is that it is also present at the bottom of Columns 1 (Tag 1, PC1) and Columns 3 (Tags 27 and 28).

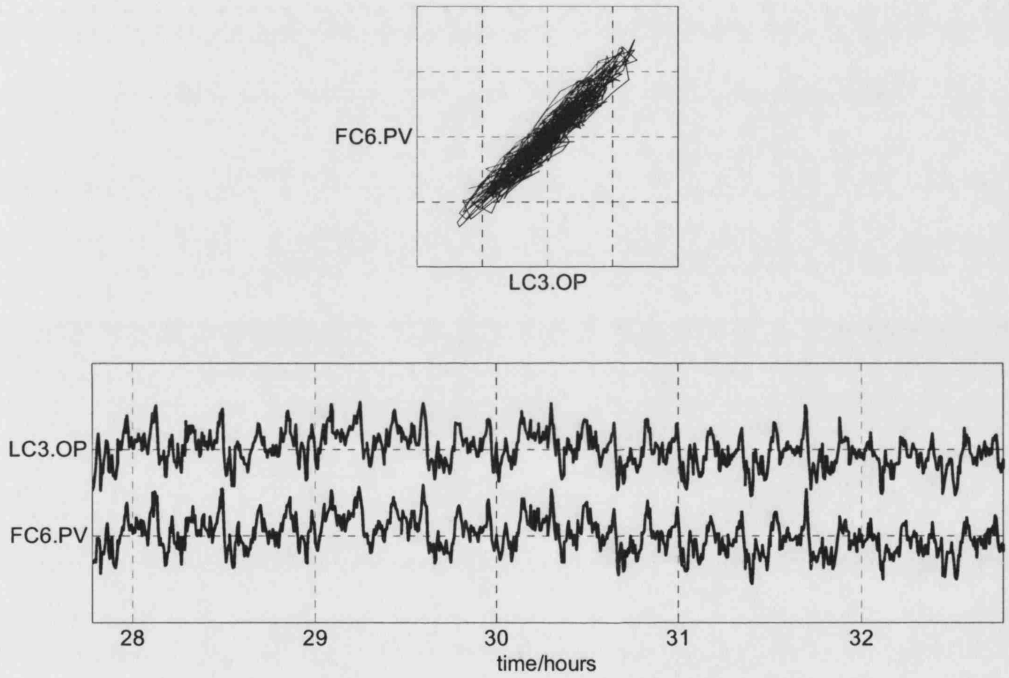

Fig 8.13 An *op-mv* plot for the LC3 (Tag 17) control loop during when the 64 samples per cycle oscillation was dominant. FC6.PV (Tag 14) is the manipulated variable.

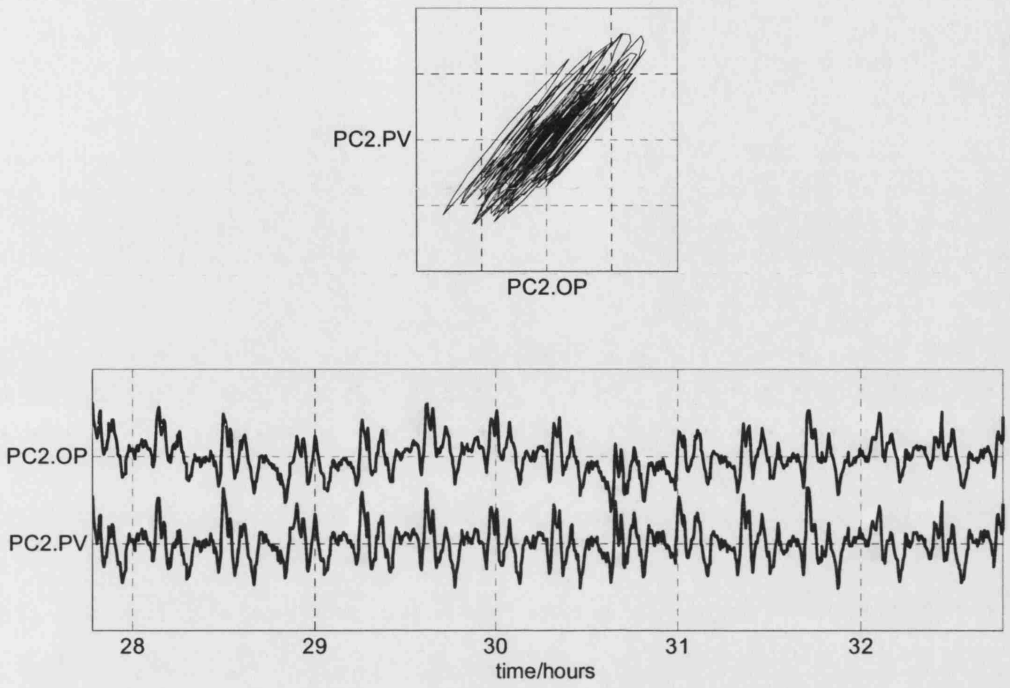

Fig 8.14 An *op-pv* plot for PC2 when the 64 samples per cycle oscillation was dominant. No *mv* measurement was available for this loop.

Root cause of the fast oscillation: The highest non-linearity in the lower panel of Figure 8.12 was in Tags 16 (PC2) and 15 (TI6). The result suggests the source of non-linearity in the 17.6 samples per cycle oscillation can also be localized to the bottom end of Column 2. No measurement is available of the flow through the PC2 control valve and, as before, the *op-pv* plot of Figure 8.15 is uninformative showing only that the controller had strong proportional action and light integral action.

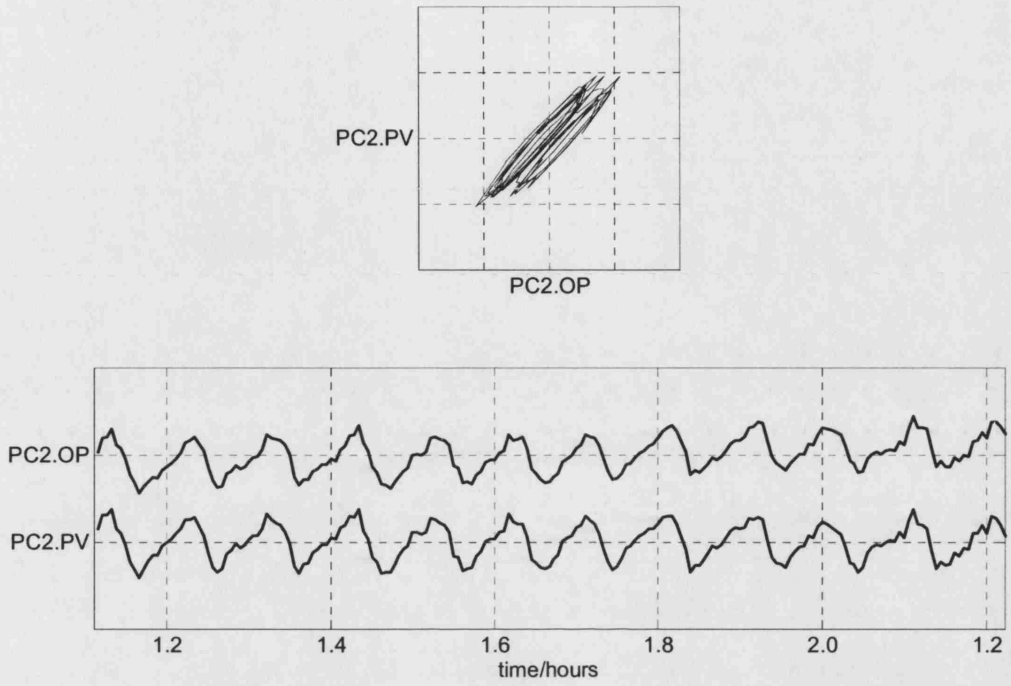

Fig 8.15 An *op-pv* plot for PC2 during an episode when the 17.6 samples per cycle oscillation was dominant.

Summary: The foregoing discussions have shown that the 64 and 17.6 samples per cycle oscillations are related because they appear sequentially in the affected tags as shown in Figure 8.11. It is possible that the disturbance itself originates in the steam system and the reason for this suggestion is that the 64 samples per cycle oscillation is present at the bottom ends of all three columns, although Columns I and 3 showed no non-linearity.

There is undoubtedly a source of the non-linearity at the bottom of Column 2, however. The steam flow valve in loop FC6 is healthy, and the more likely candidate is the PC2 (Tag 16) pressure control valve or a process non-linearity.

No further actions were taken on the 64 and 17.6 samples per cycle oscillation by the Eastman engineers. The reason for this decision was that these disturbances were minor compared to the major disturbance caused by the 341 samples per cycle (114 minute) slow oscillation which was much more widespread and powerful.

8.3.2 Diagnosis of the slow plant-wide oscillation

Non-linearity analysis: Non-linearity analysis was completed for the tags with the 341 samples per cycle oscillation with a period of 114 minutes. Earlier results published in Thornhill *et. al.* (2003*a*) showed Tags 22, 23, 25 and 29 to be non-linear with the highest non-linearity being in tag 29 (FI3). Those results were generated before the improved end-matching criterion was devised (see section 6.3). Twelve cycles of oscillation were used between samples 720 and 5000; the data were sub-sampled by 10 and the parameters were $E = H = 30$, $L = 1$, $k = 8$ and 50 surrogates (see Table 8.3).

The numerical values of the results with the improved end-match condition were slightly different as shown in the middle column of Table 8.5. The conclusion was the same, namely that Tags 22 (LC2) and 29 (FI3) have the highest non-linearity. The locations of those tags reveal the source of the 341 samples per cycle oscillation and direct the plant engineers towards focused further testing and maintenance activity in the control valve of level controller LC2 (Tag 22).

name	tag no	original calculation with filter of [100 to 1024] *N*	new calculation with filter of [100 to 1024] *N*	new calculation with filter of [70 to 1024] *N*
LC1	5	-	-	-
FC1	7	-	-	-
FC2	8	-	-	-
TI5	11	-	-	-
TI4	12	-	-	-
TC1	13	-	-	-
FC5	19	-	-	-
FI5	20	-	-	-
LC2	22	**1.1**	**1.3**	**1.3**
FC8	23	**1.1**	**1.1**	-
FI4	24	-	-	-
TC2	25	**1.0**	**1.0**	**1.0**
TI8	26	-	-	-
TI7	27	-	-	-
PI2	28	-	-	-
FI3	29	**1.1**	**1.9**	**2.1**

Table 8.5 Non-linearity analysis for the 341 samples per cycle oscillation in the Eastman Chemicals case study.

Discussion of pre-filtering: In non-linearity detection it is normally desirable to remove very low frequency non-steady trends. It is also desirable to retain higher frequencies if possible because higher frequency harmonics of a non-linear signal define the shape of its waveform and its predictability relative to its surrogates. In this case study, however, the presence of multiple oscillations requires the usual approach of retaining high frequencies to be modified because the

faster 64 and 17.6 samples per cycle oscillations interfere with the calculations. If other unrelated oscillations are present (i.e. oscillations that are not harmonics) they will upset the predictability on which the test is based.

The filter used was 100 to 1024 samples per cycle meaning that it removed any signal components having periods longer than 1024 samples per cycle and any with periods shorter than 100 samples per cycle. The third harmonic of the 341 samples per cycle oscillation is retained by this filter. The right hand column of Table 8.2 shows that little was gained by including the fourth harmonic at about 82 samples per cycle and the conclusion that tags 22 and 29 give the location of the source did not change.

An op-mv plot: Tag 22 is the LC2 controller for the decanter associated with Column 2. It is desirable to plot a trajectory showing the relationship between the demand to the valve (the controller output or *op*) and the flow through the valve. However, in this case the flow through the LC2 control valve is not measured.

An inspection of the process schematic suggests that Tag 29, FI3, could be used as a proxy measurement for the unmeasured flow through the LC2 control valve because it is the same flow plus the flow from the second decanter. The wanted *op-mv* plot (Figure 8.16) was therefore constructed from LC2.OP (Tag 22 in the OP plots of Figure 8.4) and FI3.PV (Tag 29).

The *op-mv* plot should be a straight line at 45° if the valve is linear. The plot in the upper panel of Figure 8.16 clearly shows the valve is not linear and that it has a deadband characteristic. The sticking of the valve can be also seen in the time trends. When the controller output changes direction the flow through the valve does not respond immediately and only starts to respond after the controller output signal has moved quite a long way.

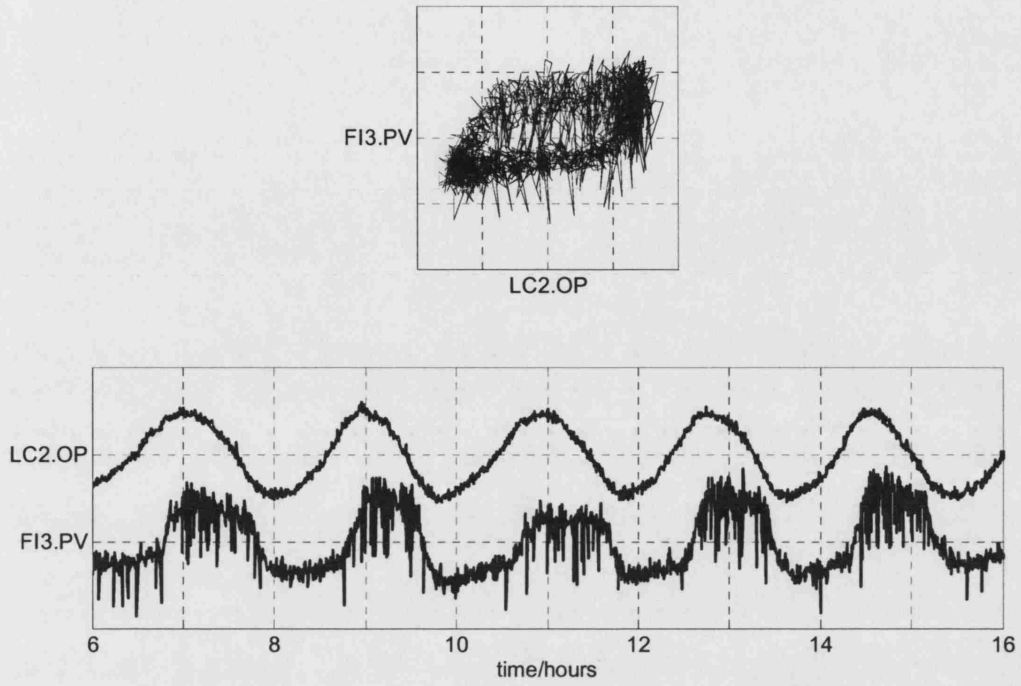

Fig 8.16 An *op-pv* plot for the LC2 (Tag 22) control valve using FI3 (Tag 29) as a proxy for the flow through the valve.

Plant testing: The Eastman engineers decided to do an interventionist plant test. They first placed the LC2 loop in manual to see if the oscillation would go away, and then they doubled the gain of the PI controller of the LC2 control loop. The results are shown in Figure 8.17.

When the controller was in manual mode from 2.5 to 4 hours the 341 samples per cycle oscillation disappeared, thus proving that the cycles originated in that control loop. Then, with the loop back in AUTO mode at 4 hours the cycles reappeared but with a different period and amplitude. The reason for this behaviour was as expected from describing function analysis and was fully discussed in section 7.3.

The control loop *op* signal is normally constant when a loop is in manual mode. In this case the movement of the LC2.OP during the period of manual operation between 2.5 and 4 hours was caused by the engineers and operators conducting valve travel tests which confirmed a deadband of 4%. The tests proved without doubt that the LC2 control valve was the source of the plant-wide disturbance and the valve was scheduled for maintenance at the next shutdown.

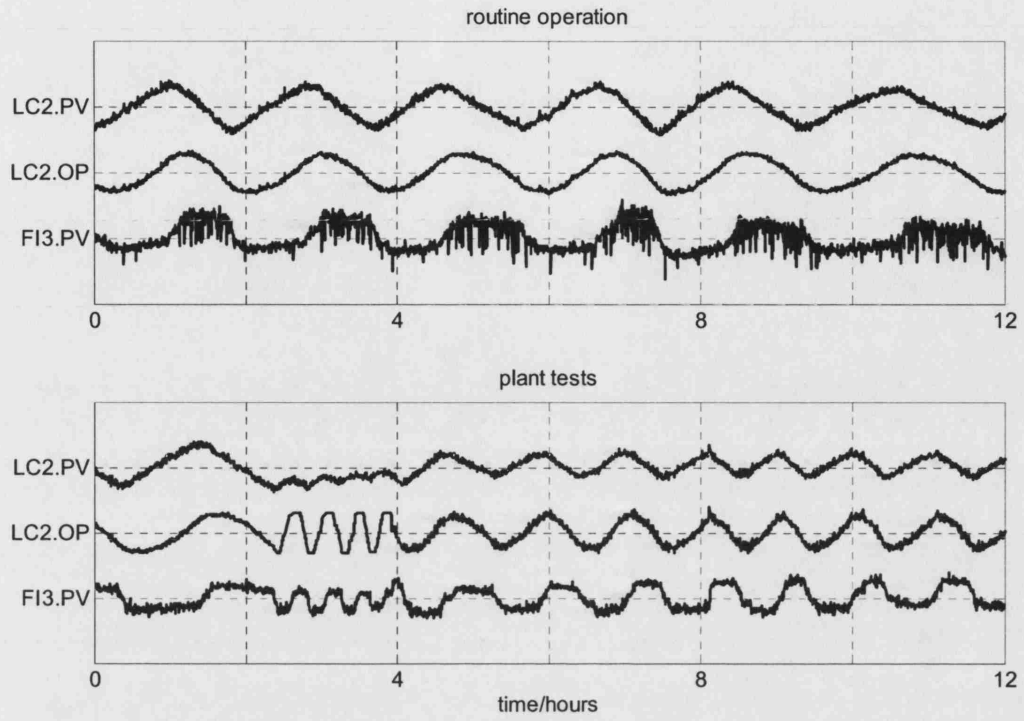

Fig 8.17 Plant tests of the LC2 control valve. Valve travel tests were conducted with the loop in manual between 2.4 and 4 hours. The loop was then put back in auto with the controller gain doubled. The LC2 valve was found to have a deadband of 4%.

8.3.3 Using the process schematic in root cause diagnosis

Using process understanding: The comments in this section will motivate future research directions in the detection and diagnosis of distributed disturbances in chemical processes. The indications are that data-driven analysis can be significantly enhanced when connectivity information from the process schematic is exploited to understand the implications of the data-driven analysis.

This subsection highlights how the process schematic and process understanding helped with the final stages of diagnosis. It thus provides an outline for a future research project to automate these steps which had to be done manually in this study.

The root cause: The data driven analysis of routine operating data yielded rich information and focused the process control engineers' attention on the part of the plant where the problem was located. The data-driven evidence pointed to LC2 (tag 22) as the root cause. At this stage it was necessary to call upon the control engineers' understanding of the plant, especially their knowledge of the connectivity of the plant.

The time trends associated with the control valve for LC2 were investigated and the need for a proxy for the unmeasured flow through the LC2 valve was identified. Examination of the flowsheet showed that FI3 (Tag 29) was equal to that flow plus material from the decanter for column 2. Therefore FI3 was used to represent the unmeasured flow through the control valve for LC2.

A future fully automated system should advise that FI3 is a proxy for the unmeasured flow through the LC2 control valve.

Mechanisms of propagation: Root cause diagnosis is not complete until a feasible mechanism of propagation to all the tags suffering from secondary oscillations is explained. In this study, as in the SE Asia refinery example in Chapter 6, the analysis was done manually by inspection of the process schematic. Mechanisms exist for the disturbance from LC2 (tag 22) to propagate to all the other tags highlighted in Figure 7.8, as follows:

- Uneven flow through the control valve of LC2 would affect FI3 (Tag 29) and propagate to column 3 including TC2 (Tag 25);
- Disturbance to TC2 would propagate to the cascade controller FC8 (Tag 23) and upset the recycle flow to column 1, and hence disturb the level controller LC1 (Tag 5).
- LC1 would adjust FC1 (Tags 7 and 8) to compensate for the disturbance to the recycle flow. It can be seen from Figure 1 that FC1 (Tags 7 and 8) and FC8 (Tag 23) are almost in anti-phase;
- It is less obvious how uneven flow through the control valve of LC2 would upset column 2 since the feed flow (FC4, Tag 10) was not affected by the plant-wide oscillation and neither was the reflux flow (FI2, Tag 18). A likely mechanism is that the feed or reflux composition vary because of disturbance to the interface level in one or both decanters;
- Disturbance to FC5 (tag 19) would also propagate to column 1 through the recycle, as described above for the FC8 recycle stream.

A future fully automated system should provide the reasoning and advise the user that a feasible path exists from the suspected root cause to all the locations suffering secondary disturbance.

8.3.4 The outcome after maintenance

Figure 8.18 shows a comparison of the behaviour of the Eastman plant before and after maintenance of the Tag 22 LC2 control valve. The plots on the right have the same vertical scales as their counterparts on the left. The valve maintenance gave a considerable reduction in plant-wide disturbance.

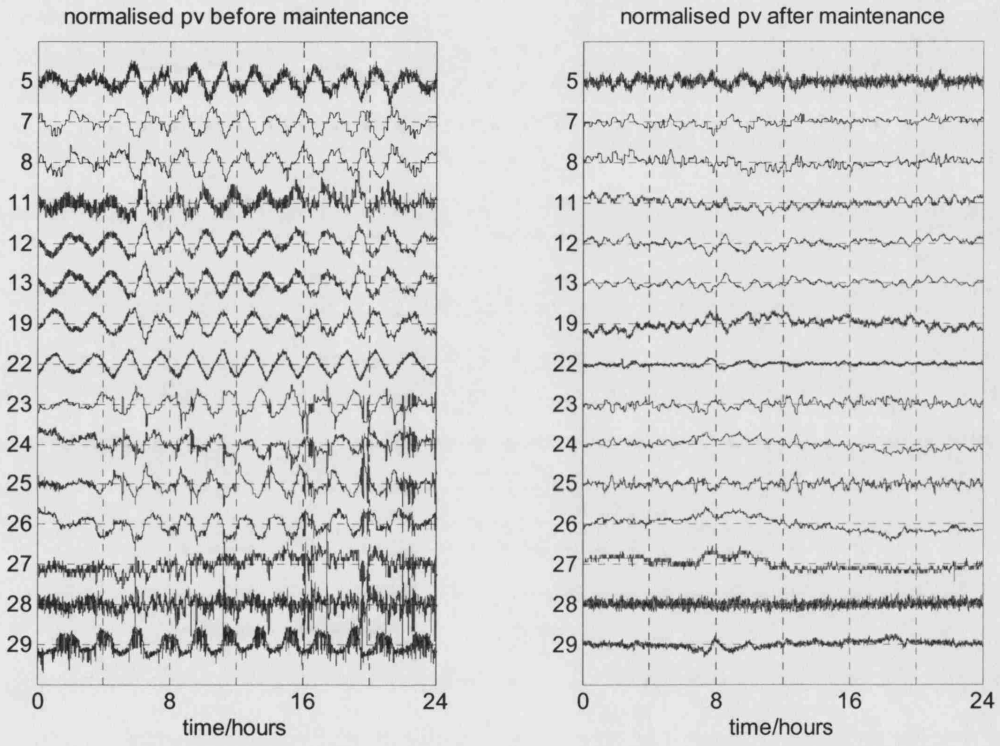

Fig 8.18 Comparison of the Eastman Chemicals process before and after maintenance of control valve LC2 (Tag 22).

8.4 CHAPTER SUMMARY

The chapter applied the methods of the thesis to an industrial case study at Eastman Chemical Company, Kingsport, Tennessee. Using only data from routine process operation, the analyses directed attention towards one particular level control valve as the root cause of a distributed disturbance with an oscillation period of 341 samples or 114 minutes. The valve was tested in situ and a deadband of 4% was confirmed. It was maintained at the next shutdown and the oscillation disappeared.

Several other distributed disturbances were also detected. Low frequency diurnal variations could be seen as well as some even lower frequency deviations. Two other oscillations at 64 and 17.6 samples per cycle were also detected and discussed. These disturbances were of lower power and affected fewer tags than the 341 samples per cycle oscillation.

The methods of the thesis worked well and the default settings suggested for the algorithms made it very easy to apply them in the new case study. The reaction of the Eastman Chemicals engineers to the results was to prioritize the disturbances in terms of the number of tags involved and the spectral power associated with the disturbance. Only one out of all the disturbances detected was considered serious enough to initiate maintenance action.

The final stages of root cause diagnosis required process understanding and a knowledge of the process flowsheet. The same requirements was observed with the SE Asia refinery data used in earlier chapters, where the presence of recycles was important. Automated linkage of cause and effect information from a process schematic with the results of data-driven analysis will be a fruitful direction for future research. Targets and performance challenges for such a system have been identified.

9. CONCLUSIONS AND RECOMMENDATIONS

9.1 OVERVIEW

9.1.1 Introduction

This chapter gives a critical review of the achievements of the PhD and highlights the conclusions and recommendations.

The subject of the thesis is the detection and diagnosis of distributed disturbances in chemical processes with practical demonstrations in real industrial processes. It uses signal processing, spectral analysis and non-linear time series analysis of measurements from routine process operations and has led to new applications of these methods in chemical process diagnosis. Industrial implementation of the methods has been considered throughout and the algorithms have been prepared carefully for technology transfer.

The chapter starts with a review of the contributions of the thesis. It also summarizes the steps taken towards technology transfer. A critical evaluation of the work of the thesis is then presented to highlight the strengths and weaknesses of the methods adopted. The issues addressed relate to the algorithms, the assumptions made in root cause diagnosis and the limitations of a purely data-driven approach.

Finally, the chapter gives recommendations for future projects. It explains the steps needed for extending the work to take account of the connectivity of the process and also considers how to extend the diagnosis methods to include a wider range of root causes.

9.1.2 Aims and outcomes of the PhD work

Thesis aims: The main aims of the PhD were outlined in Chapter 2 as follows:

- To use historical process data for plant auditing;
- To detect distributed disturbances that affect many measurements in a unit, plant or site;
- To determine which measurements are affected;
- To characterize the disturbances in order to gain insights into their severity;
- To diagnose the root cause of each disturbance;
- To consider technology transfer issues.

Dissemination: The work of the PhD has contributed to all these areas and has disseminated the findings through the articles listed below. The publication of peer-reviewed journal papers indicates that the work is novel and is of interest to academic and industrial researchers.

1. Thornhill, N.F., Shah, S.L., Huang, B., and Vishnubhotla, A., 2002, Spectral principal component analysis of dynamic process data, *Control Engineering Practice*, 10, 833-846.
2. Thornhill, N.F., Cox, J.W., and Paulonis, M., 2003*a*, Diagnosis of plant-wide oscillation through data-driven analysis and process understanding, *Control Engineering Practice*, 11, 1481-1490.
3. Thornhill, N.F., Huang, B., and Zhang, H., 2003*b*, Detection of multiple oscillations in control loops. *Journal of Process Control,* 13, 91-100.
4. Thornhill, N.F., 2005, Finding the source of a limit cycle oscillation, *IEEE Transactions on Control System Technology*, 13, 434-443.
5. Thornhill, N.F., Choudhury, M.A.A.S., and Shah, S.L., 2004, The impact of compression on data-driven process analysis, *Journal of Process Control*, 14, 389-398.
6. Choudhury, M.A.A.S., Thornhill, N.F., and Shah, S.L., 2005, Modelling of valve stiction, *Control Engineering Practice,* 13, 641-658.

The thesis author (Thornhill) wrote the text, created the figures and was responsible for the ideas and algorithms in the first four papers. The detailed contributions of the co-authors authors have been identified in the relevant chapters of the thesis. The fifth paper was written by Thornhill who also created most of the figures and the results. The compression detection algorithm in the fifth paper was a joint effort with Shoukat Choudhury at the University of Alberta, as stated in Chapter 3.

The sixth paper was mainly written by Choudhury and the thesis author contributed the text and figures for the sections on physics-based modelling of a control valve and the describing function analysis, as stated in Chapter 7.

9.2 CONTRIBUTIONS AND ACHIEVEMENTS

9.2.1 Contributions to auditing of historical plant data

Introduction: The review in Chapter 2 found that the work of the thesis is in the area termed by Venkatasubramanian *et. al.,* (2003*a*, *b*, *c*) as *quantitative process history based methods*. That is to say, the detection and diagnosis methods are driven by plant data rather than by the use of first principles models of the plant.

The techniques presented in the thesis require batches of historical plant data and then apply signal processing, spectral analysis and non-linear time series analysis to detect and diagnose

distributed disturbances, i.e. those that are present simultaneously in several places. The methods applied involve integral transforms of the process data (e.g. Fourier transform) and therefore are suitable for retrospective analysis of a batch of data. They are not designed for real-time application. Therefore the contribution of the thesis is in plant auditing using data stored in a data historian, and it falls in the same category as commercial control loop auditing products such as Process Doctor (Matrikon) Plant Triage (Expertune) and Loop Scout (Honeywell). The thesis gives a clear advance on these products, however, because of its plant-wide focus whereas the products mentioned mainly focus on the auditing of single-input-single-output loops taken one at a time.

Preparation of the historical data: The thesis has made contributions towards the preparation of historical data to get it ready for analysis. The simple concept of high density plotting has been used throughout to show data compactly. Its main benefit is that it shows time trends and spectra from one or more days of running at multiple measurement points in a chemical plant. Such a display can be very revealing but it is not a routine part of a plant operator's console. For instance, the upper panel in Figure 9.1 shows the high density plot for the Eastman Chemicals case study from Chapter 8, and also shows the plot typical of an operator's screen in the lower panel. The operator generally sees the behaviour of three or four measurements during the last 20 to 30 minutes. It would be impossible to tell that a major plant-wide oscillation is present by looking at the operator's display.

The plots can be further refined using the frequency domain filtering tools which zoom in on the behaviour in a selected frequency range. The filtering has been set up to meet the requirements of process control engineers who generally prefer to make time-domain rather than frequency domain specifications. Thus the cut-off for the filter is specified in terms of samples per cycle. For instance, visual inspection of Figure 9.1 shows the oscillation has a period of about 2 hours. The sampling interval was 20 seconds, thus the oscillation is roughly 360 samples per cycle and a filter with a pass-band between 500 samples per cycle and 100 samples per cycle would be suitable to clean up the signals. It would remove both long period (low frequency) non-stationery effects and short period (high frequency) noise.

Data compression: An important contribution of the thesis has been the examination of the impact of data compression on the results of historical plant data analysis. Even basic statistics such as mean and standard deviation are changed if data are compressed using popular methods such as the Swinging Door algorithm, and other more advanced analyses such as minimum variance control loop benchmarking (the Harris Index) also become unreliable. An algorithm for compression detection was devised with Shoukat Choudhury (University of Alberta) and a strong recommendation given to exclude any compressed data from a historical plant audit.

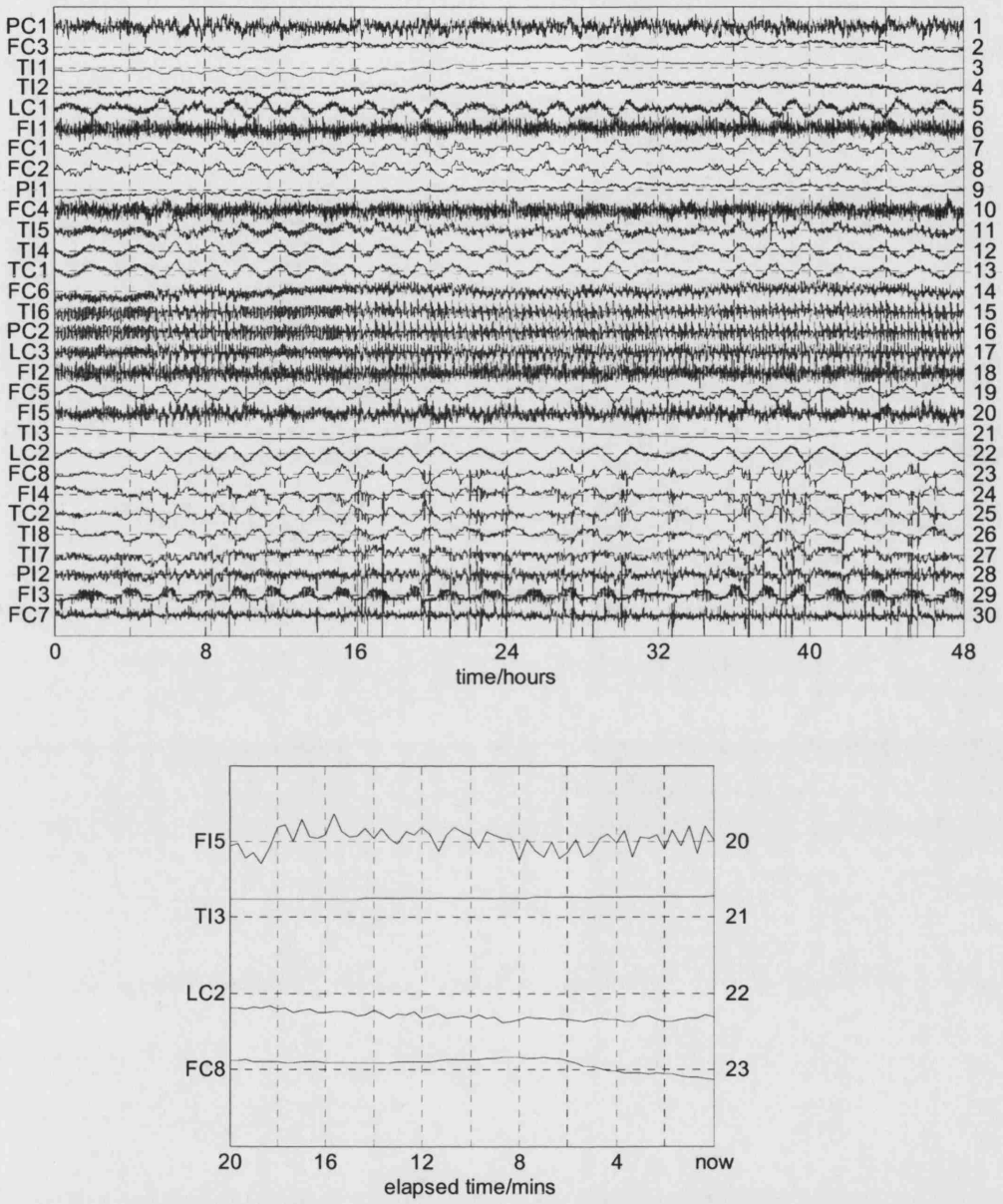

Fig 9.1 Time trends from Chapter 8: Upper panel: High density time trend plots for two days running showing a plant-wide oscillation. Lower panel: Typical operator's screen showing 20 minute trends for tags 20-23 which does not reveal the oscillation.

9.2.2 Contributions to detection and diagnosis of distributed disturbances

Distributed disturbances: A *disturbance* was defined in Chapter 6. The definition follows the minimum variance benchmarking approach to controller performance assessment of Harris, (1989) and declares that a *disturbance* is considered to be present when the time series of a plant measurement deviates from white noise. It has non-random features that can most readily be seen as non-flat features in the power spectrum. A *distributed disturbance* is a disturbance with similar characteristics present in many places in a chemical process. The similarly in the case of an oscillatory disturbance would be that its frequency of oscillation is the same, while in a more general case the measurements affected by the distributed disturbance would have similar spectral features. The thesis contributes two new methods for detection of distributed disturbances.

A key issue with distributed disturbance is to distinguish the primary disturbance (the root cause) from the secondary propagated disturbances that appear elsewhere in the plant.

Detection of distributed disturbances – Oscillation detection: Oscillations are detected using zero crossings of the autocovariance function, with a significant improvement over previous zero crossing methods which used time trends. The improvements arise for two reasons: (*a*) the autocovariance function is less noisy than the time trend and the zero crossings can therefore be identified more easily; (*b*) the calculation of the autocovariance as the inverse discrete Fourier transform of the double-sided power spectrum gives the opportunity to apply frequency domain filtering and to focus on selected frequency ranges. Multiple oscillations present in the same measurement can thus easily be detected using an automated algorithm for selection of the filters (Section 4.3.3).

A new proof was given that, for sampled data, it is more accurate to determine the period of oscillation from the intervals between the zero crossings of the autocovariance function than from the position of a spectral peak in the discrete Fourier transform (DFT). The basic reasons behind this finding are (*a*) the position of the true spectral peak is between two frequency channels except in the special case when the data window captures a complete number of cycles of oscillation. Therefore there is an error Δf in determining the frequency which translates to an error in the determination of the period of $\Delta T_p = \Delta f \times T_p^{\,2}$, which becomes larger as the period of oscillation, T_p, gets longer; (*b*) the error in determining the period of oscillation from zero crossings is insensitive to whether a complete number of cycles of oscillation have been captured. Any error is due to the fact that the zero crossings may lie between two samples, and therefore has a maximum value of $0.5 \times \Delta T$ where ΔT is the sampling interval, or less if the result is averaged over several cycles.

Detection of distributed disturbances – Spectral PCA: The method of spectral principal component analysis (Spectral PCA) detects both oscillatory and non-oscillatory disturbances by inspection of the power spectra. The most significant feature of spectral PCA is its invariance to the lags and time delays which have been extensively discussed in the literature as problems when PCA is applied to measurements from dynamic processes. The invariance property is achieved by using the power spectra determined from the squares of the absolute values of the discrete Fourier transform (DFT). Information about time delays is not lost, however. If needed, the time delay of one time trend relative to another can be determined from the phase of the DFT.

The outcome of spectral PCA is a score plot, a representation of the data in which each complete spectrum maps to a single spot. Measurements (tags) whose power spectra have similar spectral features have similar scores and are therefore close together in the score plot.

A post-processing step finds groups of tags whose spectra are sufficiently similar to form a cluster, and each cluster is interpreted as a distributed disturbance. An innovation in the clustering procedure is the use of angular separation rather than Euclidian distance to establish cluster membership. The reason for this was the observation that the clusters in spectral PCA score plots such as Figure 5.3 form plumes radiating from the origin rather than spherical clusters.

A second useful feature of the clustering algorithm is the hierarchical classification tree based on the angular distance. The tree overcomes the problem of visualization of high-dimension PCA score plots since the angles between score vectors can by as easily computed for a seven–PC model as for a two– or three–PC model.

Diagnosis of distributed disturbances: As reported by several authors, most recently by Desborough and Miller (2002), many disturbances in chemical processes are caused by faulty control valves having non-linear characteristics such as dead band and excessive static friction. The limit cycles arising from sticking control valves in a feedback control loop can propagate widely. For that reason, the work of the thesis has focused on the diagnosis of non-linear root causes. The methods developed for the diagnosis of sticking control valves are also able to locate faults originating in process non-linearities such as periodic foaming in a distillation column.

In the author's opinion, the work on non-linear root cause analysis is the most significant contribution of the thesis. The key new idea is that non-linearity of the time trends of process measurements can be used to locate the root cause of a plant-wide disturbance. Following Kantz and Schreiber (1997) a *non-linear time trend* is a signal that cannot be described as the output of a linear system driven by white noise. It is characterized by phase coherence and, if it is oscillatory, by the presence of harmonics. A measurement close to a non-linear source has more non-linearity than a measurement far away from the source; the reason for this is presumed to be that the process acts as a mechanical low-pass filter removing spectral harmonics and phase coherence as the disturbance propagates through the plant. A nonlinearity test based on a comparison of the predictability of a test time trend and its linearized surrogates was adapted for use with oscillating time trends. Two case studies involving faulty control valves have been

solved in this thesis, one of which was published as a case study with co-authors from the Eastman Chemical Company (Thornhill *et. al.*, 2003*a*). A further paper published in IEEE Transactions on Control Systems Technology (Thornhill, 2005) successfully solved further examples including the discovery of a faulty steam flow sensor and a hydrodynamic instability in a distillation column. Finally, unpublished work has shown the non-linearity method also works well for the detection of non-periodic slugging flows from oil wells.

Non-linearity detection is being actively researched by other groups, of particular note is the bispectral analysis work at the University of Alberta. A paper from that group submitted for the 2005 IFAC World Congress has also found the non-linearity as assessed by bicoherence to be highest close to the root cause in a case study at a Mitsubishi Chemicals plant in Japan. The analysis of non-linearity in process time trends is a promising area of research at present.

Physics-based understanding of the origins of disturbances: Chapter 7 of the thesis presented a first principles analysis of the physical origin of disturbances due to sticking control valves.

There were two reasons for doing the first principles modelling: (*a*) to clarify the confused literature regarding the definition of the term stiction, and (*b*) to validate an empirical model of a valve with deadband and stiction developed by Choudhury of the University of Alberta. One outcome of the work is a proposal in Choudhury *et. al.* (2005) for a new definition of stiction in line with the ANSI definitions of other non-linear behaviour such as deadband and hysteresis. The definition is given in terms of the input and output behaviour rather than in terms of the mechanics of stiction.

The first principles valve model used Newton's second law and a simple friction model which included Coulomb friction, a moving friction force that depended on velocity and static friction which exceeded the dynamic friction (Figure 7.4). Simulations based on the model showed that a valve deadband is due to Coulomb friction while stiction occurs only when the static friction exceeds the moving friction, giving rise to a jumping behaviour when motion starts. In the author's opinion, this work clarifies a the literature and is the first time the two effects have clearly been distinguished in the context of control valves in closed loop operation. The first principles model also validated Choudhury's empirical model in which he had included both a jump and a deadband whose sizes could be specified independently of one another.

A classical describing function analysis of the stiction non-linearity (Section 7.4) has confirmed results that were previously known empirically, for instance that a deadband-only non-linearity cannot cause limit cycling in a flow loop (a first order lag) with a proportional plus integral controller.

9.2.3 Technology transfer

Introduction: It was intended from the outset that the work of the PhD should be delivered in a form suitable for technology transfer. This has been achieved (*a*) by demonstrating the utility of the methods to successfully solve distributed disturbances in industrial case studies and (*b*) by selection of robust and optimized parameters for the algorithms so that they can be rolled out to new processes with a minimum of customization. The codes were also thoroughly documented, with sets of test data and test results included in the documentation.

Industrial case studies: The success of the industrial studies reported in this thesis have contributed to the technology transfer by proving the method can be readily applied to real plant-wide case studies. They have also shown a limitation of the methods that will be discussed in more detail in section 9.3.2, namely that they do not utilize information about the plant layout and connectivity.

Optimization of the algorithms: Successful technology transfer of an algorithm is crucially dependent on the choice of robust and optimized parameters. It is infeasible to expect end-users to specify settings each time the disturbance detection and diagnosis tools are applied; if that were a requirement then the users would reject the tools as too awkward to use.

Several automated algorithms have been presented in the thesis that can be applied to new plant data without customization. Examples include the detection of compression, the automated choice of filter cut-off thresholds and the selection of the optimizing parameters for non-linearity assessment. In the latter case, recommendations have been made for the number of cycles of oscillation and the number of samples per oscillation. The data from each new case is then adjusted by truncation and sub-sampling to lie within the specified bounds, a procedure that is readily automated once the period of oscillation has been determined. All that is needed is for the user to select which disturbance is to be analyzed and the previously-calculated oscillation period is then used to determine the parameters.

ABB's PDA tools: It can be concluded that technology transfer has been successful because ABB has started development on a Plant Disturbance Analysis software (PDA). PDA uses the algorithms from this thesis and will be offered to customers with ABB's Loop Performance Manager as part of their Optimize^{IT} suite. Screen shots from the new graphical user interface created by ABB are presented in Figure 9.2 which shows oscillation cluster analysis and a spectral PCA hierarchical tree for the SE Asia data set used to illustrate Chapters 3 to 6. The tree looks different from the one shown in Figure 5.8 because of enhancements suggested by Hallgeir Melbø of ABB Corporate Research in Oslo. It shows the angular separation, θ, on the vertical axis rather than $1-\cos(\theta)$ and only uses the spectra with the largest scores to construct the tree. These enhancements have improved the visual impact and interpretation of the tree.

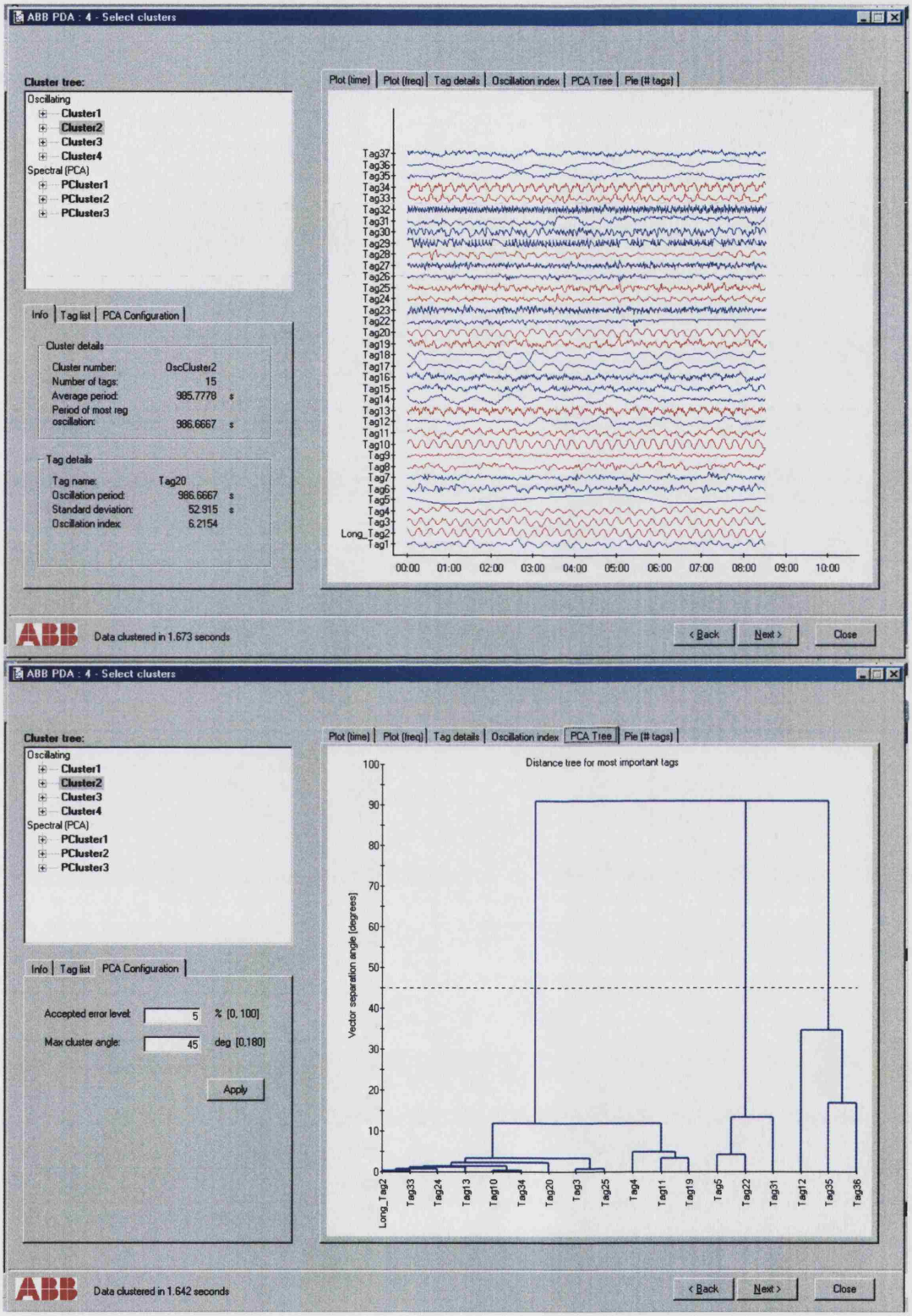

Fig 9.2 Screen shots from ABB's PDA tool (version β2.1) which uses the methods of the thesis.

9.3 CRITICAL EVALUATION

9.3.1 Evaluation of algorithms

Introduction: The algorithms presented in the thesis have all been peer reviewed and published. Some feedback is becoming available, and also some further developments have taken place in collaboration with other groups. The main issues are presented in this sub-section.

Spectral PCA: An important feature of spectral PCA is its invariance to lags and delays caused by process dynamics. This strength has been recognized by reviewers and in informal feedback from other researchers. A drawback of the method is that it can only be used for historical auditing because the DFT which is an integral transform requires the complete data set to be already in place.

Some improvements to the algorithms are needed when a data set is not stationery. The Eastman Chemicals study showed an example where an oscillation was present in only part of the data set (Section 8.3.1). Spectral PCA and oscillation analysis detected its presence in several tags but a visual inspection of the high density plot was needed to discover that the oscillation was intermittent. Intermittent behaviour might, however, be crucial to the diagnosis, for instance when the oscillation is present only under certain operating conditions. A wavelet pre-processing step to give a time-resolved frequency analysis would be helpful in the identification of such effects.

Sometimes the spectral PCA decomposition is hard to interpret because the basis functions from which the spectra in the data set are reconstructed contain multiple spectral peaks. The complication does not affect the detection of clusters in the score plots but it does make their interpretation less straightforward. Spectral independent component analysis (spectral ICA) outlined in Chapter 2 has been shown to give an improvement over spectral PCA because the basis functions in spectral ICA generally contain just one spectral peak. The theoretical reason for this improved behaviour has been analyzed by Xia (2003). Spectral ICA is a post-processing step following from spectral PCA and has been used to extend the work of the thesis by Xia *et. al.* (2005) using the data set of Chapter 8. ICA enhanced the analysis and interpretation of which tags were participating in which disturbance.

Non-linearity detection: The non-linearity detection algorithm used in the thesis is from a class of methods using embedding and surrogate data. It is intuitively simple and easy to explain, however it is an indirect method for detecting the phase coherence that characterizes a non-linear time series. The algorithm only requires about 12 cycles of oscillation sampled at roughly 20 samples per cycle (a total of 240 samples) to generate a result and is therefore ideal for short data

sets or for cases where oscillations are intermittent. Indeed, long sequences cannot be analyzed because the algorithm execution time increases proportional to N^2 where N is the number of samples.

It remains an open question whether the method presented in this thesis is a better approach than other non-linearity detection methods, especially the bicoherence method based on the bispectrum that has also been recently applied to chemical process data. Bicoherence can deal with and, in fact, it requires a long data sequence. It is somewhat complicated to explain and to use the algorithms correctly which reduces its appeal. On the other hand, it gives a more direct indication of phase coherence than the method used in this thesis. A comparative study using the same data set would be a useful next step.

9.3.2 Limitations of the work of the thesis

The need for information on process connectivity: The case study of Chapter 8 was successfully solved by the tools but it also highlighted a limitation of the data-driven approach because there is no information in the data-driven detection and diagnosis procedure about the physical relationships between the measurements. If information about the plant equipment and its connectivity were combined with the results of data-driven analysis then the diagnosis step could become significantly more automated because, for instance, an upstream valve could be identified as the more likely root causes of a propagated disturbance if there is more than one candidate. It would also be beneficial to have an automated procedure to check that there is a feasible propagation path from a candidate root cause to all other measurement points where secondary disturbances have been detected. At present a process control engineer has to do a visual inspection of the process schematic to achieve these things.

It would be useful to have an automated method for representation of and for automated reasoning using the following information:

- Connections between items of plant equipment;
- Directions of flow of mass and heat, and identification of any recycles;
- Whether a measurement is a stand-alone indicator or in a feedback control loop;
- Proxy measurements that might substitute for unmeasured quantities

Many researchers have attempted to create automated reasoning systems such as digraphs for chemical processes (see section 2.2.2). The general impression is that there are many options for representation and automated reasoning but the main stumbling block is always the capture of the process connectivity information from the schematic. In the author's opinion this is where attention should be focused in order to achieve a significant step forward in fault diagnosis.

Addressing other types of root causes: The work of the thesis has solved the problem of finding the root cause of a plant-wide disturbance originating in equipment or process non-linearity (№ 1 in the list below). It is a useful step forward because reports in the literature suggest that non-linear root causes especially sticking valves are common and widespread in chemical processes. There are, however, several other causes of distributed disturbances given by 2 to 5 in the list below:

1. Non-linearities such as saturation, dead band, or hysteresis in control valves, sensors or in the process itself that cause limit cycle oscillations;
2. Control loop interactions arising when two controllers have a shared mass and/or energy store (e.g. pressure and level controllers may compete for control of the contents of a reactor);
3. Structural disturbances caused by coordinated transfers of mass and/or energy between different process units, especially when a recycle is present;
4. Disturbances entering at plant boundaries (e.g. due to site utilities such as steam);
5. Poorly tuned controllers.

Plant-wide distributed disturbances can originate from any of the above root causes. Control loop interactions (2) are common in process systems and in the author's experience are the next most common cause after non-linearities. The focus of academic activity has been on the design and tuning of decentralized control structures to avoid interactions but not on the detection, isolation and diagnosis of disturbances caused by interaction. Structural disturbances of type (3) have been observed in processes with recycles and are becoming more prevalent as the industry moves towards increasingly integrated processes. The detection of poorly tuned loops SISO loops (5) is routine using commercial tools such as Plant Triage (Expertune), Loop Scout (Honeywell) and ProcessDoctor (Matrikon). ProcessDoctor also gives diagnosis of poor tuning but does not consider propagated secondary effects.

Advanced signal processing methods offer some possible approaches to the diagnosis of these faults. Examination of cases where structural disturbances and interacting controllers are known to be the cause suggests that the methods of normal mode analysis might be useful. The reason for the suggestion is that interacting controller often have a characteristic amplitude modulated pattern that originates from the interference of two coupled modes of oscillation having similar frequencies. The modulation can be easily detected but the challenge will be to find characteristics of these signals that uniquely identify which are the interacting controllers and also to distinguish them from other controller that are not interacting but which are upset by the disturbance.

Disturbances originating at plant boundaries might most easily be diagnosed by observing that the maximum strength (signal power) in the disturbance is located where feeds or site utilities enter the plant. One of the disturbances in Chapter 8 was identified as possibly being due to site steam because it was present at the reboiler end of more than one distillation column. Therefore the automated reasoning using plant connectivity information will help with the identification of disturbances at plant boundaries.

9.4 CHAPTER SUMMARY

This chapter has reviewed and given a critical evaluation of the work presented in the thesis.

The topic of the thesis is a new plant-wide approach to fault diagnosis which addresses the fact that disturbances originating from a root cause propagate through the process and cause secondary upsets with similar characteristics. The methods and algorithms of the thesis have been taken up by a major technology vendor and will be supplied to end users as a Plant Disturbance Analysis tool.

The strengths and weaknesses of the work were discussed in this chapter. The PhD used quantitative data-driven process history methods and its strengths arise from novel applications of signal processing, spectral analysis and non-linear time series analysis to measurements from routine process operations. The contributions to chemical process monitoring and main achievements are as follows:

- Detection and characterization of oscillatory and non-oscillatory distributed disturbances, including cases where multiple disturbances are present simultaneously;
- Distinguishing between the primary root cause disturbance and secondary propagated disturbances;
- Fundamental understanding of the mechanism and symptoms of stiction in control valves;
- Dissemination of the work through peer reviewed journals;
- Discovery of the root causes of plant-wide disturbances in industrial case studies;
- Provision of robust, easily configured algorithms for technology transfer.

Algorithmic and theoretical contributions include:

- Observation that non-linearity in the time trends of plant measurements is greatest in those measurements closest to the root cause;
- Making principal component analysis (PCA) invariant to the lags and time delays of dynamic process data, by means of spectral PCA;
- Enabling PCA clusters to be visualized when more the three principal components are used, by means of a hierarchical tree based on angular separation.

Critical evaluation has highlighted some suggestions for further work. The main criticism arises from a generic limitation of data-driven methods, which is that process knowledge and insights have to be supplied by the user. It would be very desirable to link results from the data-driven methods with information about the process. For instance, the user always needs to know the location in the plant of the disturbed measurements and the relationships between the measurements.

Recommendations for future work are:

- Linkage of the data-driven methods of the thesis with process connectivity information captured from the process flow sheet, for more advanced and automated root cause analysis;
- Diagnosis of other root causes such as controller interaction and structural oscillations;
- Improvement of the algorithms by means of:
  - detection of non-stationery behaviour, for instance by use of wavelet analysis;
  - post-processing of spectral PCA results with spectral ICA to enhance interpretation of results;
  - comparison of the bicoherence method of non-linearity detection with the method used here.

REFERENCES

Albers, J.E., 1994, Data reconciliation with unmeasured variables, *Hydrocarbon Processing (March)*, 73, 65-66.

Alwan, L.C., Champ, C.W., and Maragah.H. D., 1994, Study of average run lengths for supplementary run rules in the presence of autocorrelation, *Communications in Statistics-Simulation and Computation.*, 23, 373-391.

Armstrong-Hélouvry, B., Dupont, P., and De Wit, C. C., 1994, A survey of models, analysis tools and compensation methods for the control of machines with friction, *Automatica*, 30, 1083–1138.

Årzén, K-E., 1994, Grafcet for intelligent supervisory control applications, *Automatica*, 30, 1513-1526

Årzén, K-E., Wallén, A., and Petti, T.F., 1995, Model-based diagnosis---state transition events and constraint equations. In Tzafestas and Verbruggen, Eds., *Artificial Intelligence in Industrial Decision Making, Control and Automation.* Kluwer Academic Publishers.

Aspentech, 2001, Analysis of data storage technologies for the management of real-time process manufacturing data, Retrieved Jul 19th 2003, from http://www.advanced-energy.com/Upload/symphony_wp_infoplus.pdf.

Åström, K. J., 1991, Assessment of achievable performance of simple feedback loops. *International Journal of Adaptive Control and Signal Processing*, 5, 3-19.

Aubrun, C., Robert, M., and Cecchin, T., 1995, Fault detection in control loops, *Control Engineering Practice*, 3, 1441–1446.

Bakshi, B.R., 1998, Multiscale PCA with application to multivariate statistical process monitoring, *AIChE Journal*, 44, 1596-1610.

Bakshi, B.R., and Stephanopoulos, G., 1994, Representation of process trends - IV. Induction of real-time patterns from operating data for diagnosis and supervisory control, *Computers and Chemical Engineering*, 18, 302-332.

Bakshi, B.R., and Utojo, U., 1999, A common framework for the unification of neural, chemometric and statistical modeling methods, *Analytica Chimica Acta*, 384, 227-247.

Basseville, M., 1988, Detecting changes in signals and systems - A survey, *Automatica*, 24, 309-326.

Basseville, M., and Benveniste, A., 1983, Sequential detection of abrupt changes in spectral characteristics of digital signals, *IEEE Transactions*, IT-29, 709-724.

Bauer, M., Thornhill, N.F., and Meaburn, A, 2004, Specifying the directionality of fault propagation paths using transfer entropy, *DYCOPS4 conference*, Boston, July 1-4 2004.

Belanger P.W., and W.L. Luyben, 1997, Inventory control in processes with recycle, *Industrial & Engineering Chemistry Research*, 36, 706-716.

Belchamber, R.M., and Collins, M.P., 1993, *Method for Monitoring Acoustic Emissions*, European Patent Office, Publication Nº 0 317 322 B1.

Bialkowski, W.L., 1992, Dreams vs. reality: A view from both sides of the gap. *Proceedings of Control Systems 92, Whistler, BC, Canada.* 283-29

Blue Circle Industries, plc., 1990, *Real Time Process Control: Improved Efficiency. Expert System Opportunities, Case Study 1*, HMSO, London.

Bristol, E.H., 1990, Swinging door trending: Adaptive trend recording? *ISA National Conference Proceedinds*, 749-753.

Casdagli, M.C., Iasemidis, L.D., Sackellares, J.C., Roper, S.N., Gilmore, R.L., and Savit, R.S., 1996, Characterizing nonlinearity in invasive EEG recordings from temporal lobe epilepsy, *Physica D*, 99, 381-399.

Champ, C.W., and Woodall, W.H., 1987, Exact results for Shewhart control charts with supplementary runs rules, *Technometrics*, 29, 393-401.

Chatfield, C., and Collins, A.J., 1980, *Introduction to Multivariate Analysis*, Chapman and Hall, London, UK.

Chen, G., McAvoy, T.J., and Piovoso, M.J., 1998, A multivariate statistical controller for on-line quality improvement, *Journal of Process Control*, 8, 139-149.

Chen, J., and Howell, J., 2001, A self-validating control system based approach to plant fault detection and diagnosis, *Computers and Chemical Engineering*, 25, 337–358.

Cheung, J.T-Y., and Stephanopoulos, G., 1990, Representation of process trends - part I. A formal representation framework, *Computers and Chemical Engineering*, 14, 495-510.

Chiang, L.H., and Braatz, R.D., 2003, Process monitoring using causal map and multivariate statistics: fault detection and identification. *Chemometrics and Intelligent Laboratory Systems*, 65, 159-178.

Chiang, L.H., Russell, E., and Braatz, R.D., 2001, *Fault Detection and Diagnosis in Industrial Systems*, Springer-Verlag.

Choudhury, M.A.A.S., 2004, *Detection and Diagnosis of Control Loop Nonlinearities Using Higher Order Statistics*, PhD thesis, University of Alberta.

Choudhury, M.A.A.S., Shah, S.L., and Thornhill, N.F., 2002, Detection and diagnosis of system nonlinearities using higher order statistics, *15th IFAC World Congress*, Barcelona, Spain.

Choudhury, M.A.A.S., Shah, S.L., and Thornhill, N.F., 2004, Diagnosis of poor control loop performance using higher order statistics, *Automatica*, 40, 1719–1728.

Choudhury, M.A.A.S., Thornhill, N.F., and Shah, S.L., 2005, Modelling of valve stiction, *Control Engineering Practice*, 13, 641-658.

Chow, E.Y., and Willsky, A.S., 1984, Analytical redundancy and the design of robust failure-detection systems, *IEEE Transactions on Automatic Control*, 29, 603-614.

Clarke, D.W., 2000, Intelligent instrumentation, *Transactions of the Institute of Measurement and Control*, 22, 3-27.

Comon, P., 1994, Independent component analysis, a new concept in signal processing, *Signal Processing*, 36, 233-287.

Conlin, A.K., Martin, E.B., and Morris, A.J., 2000, Confidence limits for contribution plots, *Journal of Chemometrics*, 14, 725-736.

Cook, P.A., 1986, *Nonlinear Dynamical Systems*, Prentice-Hall International, London.

Crowe, C.M., 1988, Recursive identification of gross errors in linear data reconciliation, *AIChE Journal*, 34, 541-550.

Crowe, C.M., 1996, Data reconciliation - progress and challenges, *Journal of Process Control*, 6, 89-98.

Davies, M., (Ed), 2003, Special issue: blind signal separation, *International Journal of Adaptive Control and Signal Processing*, 18(3).

Desborough, L. and Harris, T., 1992, Performance assessment measures for univariate feedback control. *Canadian Journal of Chemical Engineering*, 70, 1186-1197.

Desborough, L., and Miller, R., 2002, Increasing customer value of industrial control performance monitoring – Honeywell's experience, in *AIChE Symposium Series* No 326, 98, 153-186.

Elsayed, E.A., 2000, Perspectives and challenges for research in quality and reliability engineering, *Int. Journal of Production Research*, 38, 1953-1976.

Emara-Shabaik, H.E., Bomberger, J., and Seborg, D.E., 1996, Cumulant/bispectrum model structure identification applied to a pH neutralization process, *IEE Proceedings of UKACC International Conference on Control 96*, UK, 1046-1051.

Ender, D.B., 1993, Process control performance: Not as good as you think, *Control Engineering* (*Sept*), 180-190.

EnTech, .1998, EnTech control valve dynamic specification (version 3.0).

Expertune Inc, 2004, *Plant Triage*, Retrieved Feb 15th 2004 from http://www.expertune.com/planttriage.html

Famili, A., Shen, W-M., Weber, R., and Simoudis, E., 1997, Data pre-processing and intelligent data analysis, *Intelligent Data Analysis*, 1, No. 1, http://www.elsevier.com/locate/ida.

Fisher-Rosemount, 1999, *Control Valve Handbook*, Marshalltown, IW, USA: Fisher Controls International Inc.

Fitzgerald, B., 1995, *Control Valves for the Chemical Process Industries,* McGraw-Hill, Inc. New York.

Forsman, K., and Stattin, A., 1999, A new criterion for detecting oscillations in control loops, *European Control Conference*, Karlsruhe, Germany.

Frank, P.M., 1990, Fault-diagnosis in dynamic-systems using analytical and knowledge-based redundancy - A survey and some new results, *Automatica*, 26, 459-474.

Frank, P.M., Ding, S.X., and Marcu, T., 2000, Model-based fault diagnosis in technical processes, *Transactions of the Institute of Measurement and Control*, 22, 57-101.

Freund, J.E., 1971, *Mathematical Statistics* (2nd ed), Prentice Hall, NJ.

Gerry, J., and Ruel, M., 2001, How to measure and combat valve stiction online, ISA, Houston, TX, USA. http://www.expertune.com/articles/isa2001/StictionMR.htm

Gertler J, and Singer D., 1990, A new structural framework for parity equation-based failure-detection and isolation, *Automatica*, 26, 381-388.

Gertler, J., 1998, *Fault Detection and Diagnosis in Engineering Systems*, Marcel Dekker.

Gertler, J., Li, W.H., Huang, Y.B., and McAvoy, T., 1999, Isolation enhanced principal component analysis, *AIChE Journal*, 45, 323-334.

Glassey, J., Ignova, M., Ward, A.C., Montague, G.A., and Morris, A.J., 1997, Bioprocess supervision: Neural networks and knowledge based systems, *Journal of Biotechnology*, 52, 201-205.

Gollmer, K., and Posten, C., 1996, Supervision of bioprocesses using a dynamic time warping algorithm, *Control Engineering Practice*, 4, 1287-1295.

Goulding, P.R., Lennox, B., Sandoz, D.J., Smith, K.J., and Marjanovic, O., 2000, Fault detection in continuous processes using multivariate statistical methods, *International Journal of Systems Science*, 31, 1459-1471.

Graebe. S.F., Goodwin, G.C., and Elsley, G., 1995, Control design and implementation in continuous steel casting, *IEEE Control Systems Magazine*, 15(4), 64-71.

Hägglund, T., 1995, A control-loop performance monitor. *Control Engineering Practice*, 3, 1543-1551..

Hale, J.C., and Sellars, H.L., 1981, Historical data recording for process computers, *Chemical Engineering Progress* (*Nov*), 38-43.

Harris, T.J., 1989, Assessment of control loop performance, *Canadian Journal of Chemical Engineering*, 67, 856-861.

Harris, T.J., Seppala, C.T., Jofreit, P.J., and Surgenor, B.W. 1996, Plant-wide feedback control performance assessment using an expert system framework, *Control Engineering Practice*, 9, 1297-1303.

Hegger, R., Kantz, H., and Schreiber, T., 2000, *TISEAN 2.1 Surrogates Manual, Periodicity Artefacts*, Retrieved Feb 15th 2004 from http://www.mpipks-dresden.mpg.de/~tisean/TISEAN_2.1/index.html

Henry M.P., 2001*a*, On-line compensation in a digital Coriolis mass flow meter, *Flow Measurement and Instrumentation*, 12, 147-161.

Henry, M.P., 2001*b*, Self-Validating (SEVA) sensors - towards standards and products, *Automazione e Strumentazione,* downloaded 25th August 2004 from http://www.trautomation.com/automation/news.nsf/02D6A1D18C7F6F114C1256A16003EABF3?OpenDocument

Henry, MP and Clarke, D.W., 2004, *Web Valve*, downloaded 25th August 2004 from http://seva.eng.ox.ac.uk/utcpage/research/webvalve.html

Hinich, M. J., 1982, Testing for Gaussianity and linearity of a stationary time series. *Journal of Time Series Analysis,* 3, 169–176.

Holly, W., Cook, R and Crowe, C.M., 1989, Reconciliation of mass flow rate measurements in a chemical extraction plant, *Canadian Journal of Chemical Engineering*, 67, 595-601.

Horch, A.,1999, A simple method for detection of stiction in control valves. *Control Engineering Practice*, 7, 1221-1231.

Horch, A., 2000, *Condition Monitoring of Control Loops*, PhD Thesis, KTH Royal Institute of Technology, Stockholm.

Horch, A., 2002, Patents WO0239201and US2004/0078168.

Horch, A., Isaksson, A.J., and Forsman, K., 2000, Diagnosis and characterization of oscillations in process control loops. *Proceedings of Control Systems 2000*, Victoria, Canada, 161–165.

Huang, B., Thornhill, N.F., Shah, S.L., and Shook, D., 2002, Path analysis for process troubleshooting, *Proceedings of AdConIP 2002,* Kumamoto, Japan, 149-154.

Hyvarinen, A., and Oja, E., 2000, Independent component analysis: algorithms and applications, *Neural Networks*, 13, 411-430.

Ignova, M., Paul, G.C., Glassey, J., Ward, A.C., Montague, G., Thomas, C.R., and Karim, M.N., 1996, Towards intelligent process supervision - industrial penicillin fermentation case study, *Computers and Chemical Engineering*, 20, S545-S550.

Iordache, C., Mah, R.S.H., and Tamhane, A.C., 1985, Performance studies of the measurement test for detection of gross errors in process data, *AIChE Journal*, 31, 1187-1201.

ISA Subcommittee SP75.05, 1979, *Process Instrumentation Terminology. Technical Report ANSI/ISA-S51.1-1979*, Instrument Society of America.

Isermann R., 1984, Process fault-detection based on modeling and estimation methods - A survey, *Automatica*, 20, 387-404.

Jackson, J.E., and Mudholkar, G.S., 1979, Control procedures for residuals associated with principal components analysis, *Technometrics*, 21, 341-349.

Jia, F., Martin, A.B., and Morris, A.J., 2000, Non-linear principal components analysis with application to process fault detection, *International Journal of System Science*, 31, 1473-1487.

Johnson, R.A., and Wichern, D.W., 1992, *Applied Multivariate Statistical Analysis*, Prentice Hall.

Jongenelen, E.M., den Heijer, C., and van Zee, G.A., 1988, Detection of gross errors in process data using Studentised residuals, *Computers and Chemical Engineering*, 12, 845-847.

Kano, M., Tanaka, S., Hasebe, S., Hashimoto, I., and Ohno, H., 2003, Monitoring independent components for fault detection, *AIChE Journal*, 49, 969-976.

Kano, M., Maruta, H., Kugemoto, H., and Shimizu, K., 2004, Practical model and detection algorithm for valve stiction, *Proceedings of the Seventh IFAC-DYCOPS Symposium*, Boston, USA, July 3-5.

Kantz, H., and Schreiber, T., 1997, *Nonlinear Time Series Analysis* Cambridge University Press.

Karnopp, D., 1985, Computer simulation of stick-slip friction in mechanical dynamical systems, *Journal of Dynamic Systems, Measurement, and Control*, 107, 100–103.

Karstang, T.V., and Henriksen, A., 1992, Infrared spectroscopy and multivariate calibration used in quantitative analysis of additives in high density polyethylene, *Chemometrics and Intelligent Laboratory Systems*, 14, 331-339.

Kassidas, A., Taylor, P.A., and MacGregor, J.F., 1998, Off-line diagnosis of deterministic faults in continuous dynamic multivariable processes using speech recognition methods, *Journal of Process Control.*, 8, 381-393.

Kayihan, A., and Doyle III, F.J., 2000, Friction compensation for a process control valve, *Control Engineering Practice*, 8, 799–812.

Kendall, M, and Ord, J.K., 1990, *Time Series (3rd Ed)*, Edward Arnold, Sevenoaks, Kent.

Koninckx, J., 1988, *On Line Optimisation of Chemical Plants Using Steady-State Models*, PhD thesis, Department of Chemical and Nuclear Engineering, University of Maryland at College Park.

Kosanovich, K.A., and Piovoso, M.J., 1997, PCA of wavelet transformed process data for monitoring, *Intelligent Data Analysis Journal*, 1:2 *on-line at* http://www.elsevier.nl/locate/ida.

Kourti, T., and MacGregor, J.F., 1996, Control of multivariate processes, *Journal of Quality Control*, 28, 409-428.

Kramer, M.A., and Palowitch, B.L., 1987, A rule-based approach to fault diagnosis using the signed directed graph, *AIChE Journal*, 33, 1067-1078.

Kresta, J.V., MacGregor, J.F., and Marlin, T.E., 1991, Multivariate statistical monitoring of process operating performance, *Canadian Journal of Chemical Engineering*, 69, 35-47.

Ku, W.F., Storer, R.H., and Georgakis, C., 1995, Disturbance detection and isolation by dynamic principal component analysis, *Chemometrics and Intelligent Laboratory Systems*, 179-196.

Kudic, A., and Thornhill, N.F., 1996, Data compression for process monitoring, *AIChE Symposium Series*, 92, 335-338.

Kundur, P., 1993, *Power System Stability and Control,* McGraw-Hill.

Lai, Y-C., and Lerner, D., 1998, Effective scaling regime for computing the correlation dimension from chaotic time series, *Physica D*, 115, 1–18.

Lakshminarayanan, S., Shah, S.L., and Nandakumar, K., 1997, Modelling and control of multivariable processes: Dynamic PLS approach, *AIChE Journal*, 43, 2307-2322.

Learned, R.E., and Willsky, A.S., 1995, A wavelet packet approach to transient signal classification, *Applied and Computational Harmonic Analysis*, 2, 265-278.

Lee, J.M., Yoo, C.K., and Lee, I.B., 2004*a*, Statistical process monitoring with independent component analysis, *Journal of Process Control*, 14, 467-485.

Lee, J.M., Yoo, C.K., and Lee I.B., 2004*b*, Statistical monitoring of dynamic processes based on dynamic independent component analysis, *Chemical Engineering Science*, 59, 2995-3006.

Lee, D.D., and Seung, H.S., 1999, Learning the parts of objects by non-negative matrix factorization, *Nature,* 401,788-791.

Li, R.F., and Wang, X.Z., 2002, Dimension reduction of process dynamic trends using independent component analysis, *Computers and Chemical Engineering*, 26, 467-473.

Luo, R.F., Misra, M., and Himmelblau, D.M., 1999, Sensor fault detection via multiscale analysis and dynamic PCA, *Industrial and Engineering Chemistry Research*, 38, 1489-1495.

Mah, R.S.H., Tamhane, A.C., Tung, S.H. and Patel, A.N., 1995, Process trending with piecewise linear smoothing, *Computers and Chemical Engineering*, 19, 129-137.

Martin, E.B., and Morris, A.J., 1996, Non-parametric confidence bounds for process performance monitoring charts, *Journal of Process Control*, 6, 349-358.

Martin, G.D., Turpin, L.E., and Cline, R.P., 1991, Estimating control function benefits. *Hydrocarbon Processing* (*June*), 68-73.

Matrikon Inc., 2004, *Control Loop Performance – ProcessDoctor*, Retrieved Feb 15th 2004 from http://www.matrikon.com/products/processdoc/.

Matsuo, T., Tadakuma, I., and Thornhill, N.F., 2004, Diagnosis of a unit-wide disturbance caused by saturation in a manipulated variable, *IEEE APC 2004 Conference*, Vancouver.

Maurya, M.R., Rengaswamy, R., and Venkatasubramanian, V., 2003*a*, A systematic framework for the development and analysis of signed digraphs for chemical processes. 1. Algorithms and analysis, *Industrial and Engineering Chemistry Research*, 42, 4789-4810.

Maurya, M.R., Rengaswamy, R., and Venkatasubramanian, V., 2003*b*, A systematic framework for the development and analysis of signed digraphs for chemical processes. 2. Control loops and flowsheet analysis, *Industrial and Engineering Chemistry Research*, 42, 4811-4827.

McMillan, G. K., 1995, Improve control valve response, *Chemical Engineering Progress*, 91(6), 76-84.

Miao, T., and Seborg, D.E., 1999, Automatic detection of excessively oscillatory feedback control loops, *IEEE Conference on Control Applications*, Hawaii, 359-364.

Mulgrew, B., Grant, P., and Thompson, J., 2003, *Digital Signal Processing*, 2nd edition, Palgrave MacMillan.

Narashiman, S., and Mah, R.S.H., 1989, Treatment of general steady-state process models in gross error identification, *Computers and Chemical Engineering*, 13, 851-853.

Narashiman, S., Mah, R.S.H., Tamhane, A.C., Woodward, J.W., and Hale, J.C., 1988, A composite statistical test for detecting changes of steady-states, *AIChE Journal*, 1409-1418.

Nikias, C.L., and Petropulu, A., 1993, *Higher-Order Spectra: A Nonlinear Signal Processing Framework*, Prentice-Hall, Englewood Cliffs, NJ.

Olsson, H., 1996, *Control Systems With Friction*, PhD thesis, Lund Institute of Technology, Sweden.

OSI Software Inc, 2002, *PI data storage component overview*, Retrieved Jul 19th 2003, from http:// www.osisoft.com/ 270.htm

Owen, J.G., Read, D., Blekkenhorst, H., and Roche, A.A., 1996, A mill prototype for automatic monitoring of control loop performance. *Proceedings of Control Systems 96*, Halifax, Novia Scotia, 171-178.

Page, E.S., 1955, Control charts with warning lines, *Biometrika*, 42, 243-257.

Pain, H.J., 1987, *The Physics of Vibrations and Waves*, Wiley.

Patton, R., Frank, P., ad Clark, R., 2000, *Issues of Fault Diagnosis for Dynamic Systems*, Springer Verlag.

Paulonis, M.A., and Cox, J.W., 2003, A practical approach for large-scale controller performance assessment, diagnosis, and improvement, *Journal of Process Control*, 13, 155-168.

Pearson, R., 2001, Exploring process data, *Journal of Process Control*, 11, 179-194.

Petersen, J., 2000, Causal reasoning based on MFM, *Proceedings of the Conference on Cognitive Systems Engineering in Process Control (CSEPC 2000)*, Taejon, Korea, 36-43.

Petti, T.F., Klein, J., and Dhurjatu, P.S., 1990, Diagnostic model processor - using deep knowledge for process fault-diagnosis, *AICHE Journal*, 36, 565-575

Piipponen, J. (1996). Controlling processes with nonideal valves: Tuning of loops and selection of valves. *Proceedings of Control Systems 96, Halifax, Nova Scotia, Canada*, 179–186.

Plumbley, M., 2002, Conditions for nonnegative independent component analysis, *IEEE Signal Processing Letters*, 9, 177-180.

Poggio, T., and Girosi, F., 1990, Networks for approximation and learning, *Proceedings of the IEEE*, 78, 1481-1497.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T., 1986, *Numerical Recipes*, Cambridge University Press.

Proakis, J.G., and Manolakis, D.G., 1996, *Digital Signal Processing*, 3rd Edition, Prentice Hall, NJ, USA

Pryor, C., 1982, Autocovariance and power spectrum analysis. *Control Engineering* (*Oct*), 103-106.

Qin, S.J., and McAvoy, T.J., 1992, Nonlinear PLS modeling using neural networks, *Computers and Chemical Engineering*, 16, 379-391.

Qin, S.J., 1998, Control performance monitoring - a review and assessment. *Computers and Chemical Engineering*, 23 173-186.

Rao, T.S. and Gabr, M.M., 1980, A test for linearity and stationarity of time series. *Journal of Time Series Analysis*, 1, 145–158.

Rengaswamy, R., and Venkatasubramanian, V., 1995, A syntactic pattern-recognition approach for process monitoring and fault-diagnosis, *Engineering Applications of Artificial Intelligence*, 8, 35-51.

Rhodes, C., and Morari, M., 1998, Determining the model order of nonlinear input/output systems, *AIChE Journal*, 44, 151-163.

Rhodes, C., and Morari, M., 1997, The false nearest neighbors algorithm: an overview, *Computers and Chemical Engineering*, 21, S1149-S1154.

Riley, M.R., Rhiel, M., Zhou, X., Arnold, M.A., and Murhammer, D.W., 1997, Simultaneous measurement of glucose and glutamine in insect cell culture media by near infrared spectroscopy. *Biotechnology and Bioengineering*, 55, 11-15.

Rollins, D.K., and Davis, J.F., 1993, Gross error detection when variance-covariance matrices are unknown, *AIChE Journal*, 39, 1335-1341.

Ruel, M., 2001, *TOP Control Optimization List*, http://www.topcontrol.com/en/papers.htm

Ruel, M., and Gerry, J., 1998, Quebec quandary solved by Fourier transform, *Intech (August)*, 53-55.

Schreiber, T., 2000, Measuring information transfer. *Physical Review Letters*, 85, 461-464.

Schreiber, T., 1999, Interdisciplinary application of nonlinear time series methods, *Physics Reports-Review Section of Physics Letters*, 308, 2-64.

Schreiber, T., and Schmitz, A., 2000, Surrogate time series. *Physica D.*, 142, 346-382.

Seasholtz, M.B., 1999, Making money with chemometrics, *Chemometrics and Inteligent. Laboratory Systems*, 45, 55-64.

Shao, R., Jia, F., Martin, E.B., and Morris, A.J., 1999, Wavelets and non-linear principal components analysis for process monitoring, *Control Engineering Pra*ctice, 7, 865-879.

Sharif, M.A., and Grosvenor, R.I., 1998, Process plant condition monitoring and fault diagnosis, *Proceedings of the Institute of Mechanical Engineering Part E*, 212, 13–30.

Shi, R.J., and MacGregor, J.F., 2000, Modeling of dynamic systems using latent variable and subspace methods, *Journal of Chemometrics*, 14, 423-439.

Shinskey, F.G., 1996, *Process Control Systems, 4th edition,* McGraw Hill.

Shinskey, F.G., 2000, The three faces of control valves, *Control Engineering (July)*, 83.

Shunta, J.P., 1995, *Achieving world class manufacturing through process control*, Prentice-Hall, NJ.

Small, M., and Tse, C.K., 2002, Applying the method of surrogate data to cyclic time series, *Physica D*, 164, 187-201.

Stam, C.J., Pijn, J.P.M., and Pritchard, W.S., 1998, Reliable detection of nonlinearity in experimental time series with strong periodic components, *Physica D*, 112, 361-380.

Stanfelj, N., Marlin, T.E., and MacGregor, J.F., 1993, Monitoring and diagnosing process control performance: The single loop case, *Industrial and Engineering Chemistry Research*, 32, 301-314.

Stephanopoulos, G., Locher, G., Duff, M.J., Kamimura, R., and Stephanopoulos, G., 1997, Fermentation data base mining by pattern recognition, *Biotchnoogy and Bioengineering*, 53, 443-452.

Stoumbos, Z.G., Reynolds, M.R., Ryan, T.P., and Woodall, W.H., 2000, The state of statistical process control as we proceed into the 21st century, *Journal of the American Statistical Association*, 95, 992-998.

Sugihara, G., and May, R.M., 1990, Nonlinear forecasting as a way of distinguishing chaos from measurement error in time-series, *Nature*, 344, 734-741.

Tan, C.C., Thornhill, N.F., and Belchamber, R.M., 2002, Principal components analysis of spectra, with application to acoustic emissions from mechanical equipment, *Transactions of the Institute of Measurement and Control*, 24, 333-353.

Tangirala, A.K., and Shah, S.L., 2005, Non-negative matrix factorization for detection of plant-wide oscillations, submitted to *IEEE Transactions on Knowledge and Data Mining.*

Terrible, J., Shahani, G., Gagliardi C., Baade, W., Bredehoft, R., and Ralston, M., 1999, Consider using hydrogen plants to cogenerate power needs, *Hydrocarbon Processing*, 78(12), 43-53.

Tham, M.T., and Parr, A., 1994, Succeed at on-line validation and reconstruction of data, *Chemical Engineering Progress (May)*, 46-56.

Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., and Farmer, J.D., 1992, Testing for nonlinearity in time-series - the method of surrogate data, *Physica D*, 58, 77-94.

Thompson, M.L., and Kramer, M.A., 1994, Modelling chemical processes using prior knowledge and neural networks, *AIChE Journal*, 40, 1328-1340.

Thornhill, N.F., Hutchison, R.J., and Haughie, B.T., 1996*a*, A probability chart for statistical process analysis, with applications, *Journal of Process Control*, 6, 161-167.

Thornhill, N.F., Sadowski, R., Davis, J.R., Fedenczuk, P., and Knight, M.J., 1996*b*, Practical experiences in refinery control loop performance assessment, *IEE Conference Publication*, 427/1, 175-180.

Thornhill, N.F., and Hägglund, T., 1997, Detection and diagnosis of oscillation in control loops, *Control Engineering Practice*, 5, 1343-1354.

Thornhill, N.F., Oettinger, M., and Fedenczuk, P., 1998, Performance assessment and diagnosis of refinery control loops, *AIChE Symposium Series 320*, 94, 373-379.

Thornhill, N.F., Shah, S.L., and Huang, B., 2000, Detection and diagnosis of unit-wide oscillations, *PCI 2000, Glasgow.*

Thornhill, N.F., Shah, S.L., and B. Huang, B., 2001, Detection of distributed oscillations and root cause diagnosis, *Proceedings of CHEMFAS-4*, Korea, 167-172.

Thornhill, N.F., Shah, S.L., Huang, B., and Vishnubhotla, A., 2002, Spectral principal component analysis of dynamic process data, *Control Engineering Practice*, 10, 833-846.

Thornhill, N.F., Cox, J.W., and Paulonis, M., 2003*a*, Diagnosis of plant-wide oscillation through data-driven analysis and process understanding, *Control Engineering Practice*, 11, 1481-1490.

Thornhill, N.F., Huang, B., and Zhang, H., 2003*b*, Detection of multiple oscillations in control loops. *Journal of Process Control*, 13, 91-100.

Thornhill, N.F., Choudhury, M.A.A.S., and Shah, S.L., 2004, The impact of compression on data-driven process analysis, *Journal of Process Control*, 14, 389-398.

Thornhill, N.F., 2005, Finding the source of a limit cycle oscillation, *IEEE Transactions on Control System Technology*, 13, 434-443.

Timmer, J., Schwarz, U., Voss, H.U., Wardinski, I., Belloni, T., Hasinger, G., van der Klis, M., and Kurths, J., 2000, Linear and nonlinear time series analysis of the black hole candidate Cygnus X-1, *Physical Review E*, 61, 1342-1352.

Tjoa, I.B., and Biegler, L.T., 1991, Simultaneous strategies for data reconciliation and gross error detection of non-linear systems, *Computers and Chemical Engineering*, 15, 679-690.

Tong, H.W., and Crowe, C.M., 1996, Detecting persistent gross errors by sequential analysis of principal components, *Computers and Chemical Engineering*, 20, S733-S738.

Tugnait, J.K., 1982, Detection and estimation for abruptly changing systems, *Automatica*, 18, 607-615.

Turner, P., Montague, G., and Morris, J., 1996, Dynamic neural networks in non-linear predictive control (an industrial application), *Computers and Chemical Engineering*, 20, , S937-S942.

Tyler, M.L., and Morari, M., 1996, Performance monitoring of control systems using likelihood methods. *Automatica*, 32, 1145-1162.

Valle, S., Li, W.H., and Qin, S.J., 1999, Selection of the number of principal components: The variance of the reconstruction error criterion with a comparison to other methods, *Industrial and Engineering Chemistry Research*, 38, 4389-4401.

Vedam, H., and Venkatasubramanian, V., 1999, PCA-SDG based process monitoring and fault diagnosis, *Control Engineering Practice*, 7, 903-917.

Venkatasubramanian, V., 2001, Process fault detection and diagnosis: Past, present and future, *Proceedings of CHEMFAS4, Korea,* 3-15.

Venkatasubramanian, V., Rengaswamy, R., Yin, K., and Kavuri, S.N., 2003*a*, A review of process fault detection and diagnosis Part I: Quantitative model-based methods, *Computers and Chemical Engineering*, 27, 293-311.

Venkatasubramanian, V., Rengaswamy, R., Kavuri, S.N., 2003*b*, A review of process fault detection and diagnosis Part II: Quantitative model and search strategies, *Computers and Chemical Engineering*, 27, 313-326.

Venkatasubramanian, V., Rengaswamy, R., Kavuri, S.N., Yin, K., 2003*c*, A review of process fault detection and diagnosis Part III: Process history based methods, *Computers and Chemical Engineering*, 27, 327-34

Verneuil Jr, V.S., and Madron, F., 1992, Banish bad plant data, *Chemical Engineering Progress* (*Oct*), 45-51.

Watson, M.J., Liakopoulos, A., Brzakovic, D., and Georgakis, C., 1998, A practical assessment of process data compression techniques, *Industrial and Engineering Chemistry Research.*, 37, 276-274.

Welch, P.D., 1967, The use of fast Fourier transforms for the estimation of power spectra. *IEEE Transactions Audio and Electroacoustics*, AU-15, 70-73.

Westerhuis, J.A., Gurden, S.P., and Smilde, A.K., 2000, Generalized contribution plots in multivariate statistical process monitoring, *Chemometrics and Intelligent Laboratory Systems*, 51, 95-114.

Wetherill, G.B., and Brown, D.W., *Statistical Process Control*, Chapman and Hall, London, (1991).

Whalen, B.R., 1983, *Basic Instrumentation* (3rd ed.). Austin, TX, Petroleum Extension Service (PETEX).

Wise, B.M., Ricker, N.L., Veltkamp, D.F., and Kowalski, B.R., 1990, A theoretical basis for the use of principal components models for monitoring multivariate processes, *Process Control and Quality*, 1, 41-51.

Wise, B.M., and Ricker, N.L., 1992, Identification of finite impulse response models by principal components regression: Frequency response properties, *Process Control and Quality*, 4, 77-86.

Wise, B.M., and Gallagher, N.B., 1996, The process chemometrics approach to process monitoring and fault detection, *Journal of Process Control*, 6, 329-348.

Wold, S., Esbensen, K., and Geladi, P., 1987, Principal component analysis, *Chemometrics and Intelligent Laboratory Systems*, 2, 37-52.

Wu, H.D., Siegel, M., and Khosla, P., 1999, Vehicle sound signature recognition by frequency vector principal component analysis, *IEEE Transactions on Instrumentation and Measurement*, 48 1005-1009.

Xia, C., and Howell, J., 2001*a*, Controller output based, single number statistics for loop status monitoring, *CHEMFAS-4*, IFAC, 127-134.

Xia, C., and Howell, J., 2001*b*, Loop Status Statistics, *CHEMFAS-4*, IFAC, 372-376.

Xia, C. and Howell, 2003*a*, Isolating multiple sources of plant-wide oscillations via spectral independent component analysis, *Proceedings of Safeprocess 2003*, Washington, D.C., USA.

Xia, C., and Howell, J., 2003*b*, Loop status monitoring and fault localisation. *Journal of Process Control*, 13, 679-691.

Xia, C., 2003, *Control Loop Measurement Based Isolation of Faults and Disturbances in Process Plants*, PhD Thesis, University of Glasgow, 2003.

Xia, C., Howell, J., and Thornhill, N.F., 2005, Detecting and isolating multiple plant-wide oscillations via spectral independent component analysis, *Automatica, accepted for publication.*

Yang, J.C.Y., and Clarke, D.W., 1997, A self-validating thermocouple, *IEEE Transactions on Control Systems Technology*, 5, 239-253.

Yeung, K.S.Y., Hoare, M., Thornhill, N.F., Williams, T., and Vaghjiani, J.D., 1999, Near infra red spectroscopy for bioprocess monitoring and control, *Biotechnology and Bioengineering.*, 63, 684-693.

Yoon, S., and MacGregor, J.F., 2000, Statistical and causal model-based approaches to fault detection and isolation, *AIChE Journal*, 46, 1813-1824.

Zang, X., and Howell, J., 2003, Discrimination between bad turning and non-linearity induced oscillations through bispectral analysis, *Proceedings of SICE Annual Conference*, Fukui, Japan.

Zang, X., and J. Howell, 2004, Correlation dimension and Lyapunov exponents based isolation of plant-wide oscillations, *DYCOPS 7*, Boston July 5-7.

Zhang, J., 2001, Developing robust neural network models by using both dynamic and static process operating data, *Industrial and Engineering Chemistry Research*, 40, 234-241.

APPENDICES

A1 THE DISCRETE FOURIER TRANSFORM (DFT)

A1.1 Basics of the DFT

Introduction: Appendix A1 reviews the features and properties of the discrete Fourier transform (DFT). It clarifies the notations used in the thesis to label the time-domain and frequency domain axes and discusses the issue of spectral leakage raised in Chapter 6. This section is standard text book material and not novel. Compared to a signal processing text book, however, there is an emphasis on the time domain representation because chemical process control engineers tend to find time domain descriptions more useful than the frequency domain. For instance, it is more intuitive and useful to characterize the period of an oscillation in terms of samples per cycle than as a frequency in Hz.

Sampling and axis labels: Axes with several equivalent axis labels are shown below. Key features are:

- The data sequence has N samples numbered from 0 to $N-1$. They may be presented on a sample number axis or on a time axis. If the sampling interval is T_S then the total duration of the sequence in time units is $(N-1)T_S$.
- There are N frequency channels also numbered from 0 to $N-1$. They may be presented on a frequency channel axis. A normalized frequency axis given by f/f_S is also useful, where f_S is the sampling frequency;
- The sampling frequency is $f_S = 1/T_S$;
- The spacing between the frequency channels is $\Delta f = 1/NT_S$;
- The frequency of the $k = N-1$ channel is $\frac{N-1}{N} f_S$. The $k = N$ frequency channel (not shown in the figures below) represents the sampling frequency f_S.

The normalized frequency axis: A benefit of the normalized frequency axis is that it can help with the description of spectral features. For instance, if the discrete Fourier transform (DFT) has a spectral feature at $f/f_S = 0.1$ it means the time domain has a sinusoidal feature whose frequency is one tenth of the sampling frequency. This can be stated in the time domain as a sinusoidal feature having ten samples per cycle. A sinusoidal signal with a large period of oscillation (say, the period is $100T_S$) maps to $f/f_S = 0.01$ on the frequency axis meaning that

the time domain feature has 100 samples per cycle. The point $f/f_s = 0.5$ on the normalized frequency axis is the Nyquist sampling frequency corresponding to two samples per cycle.

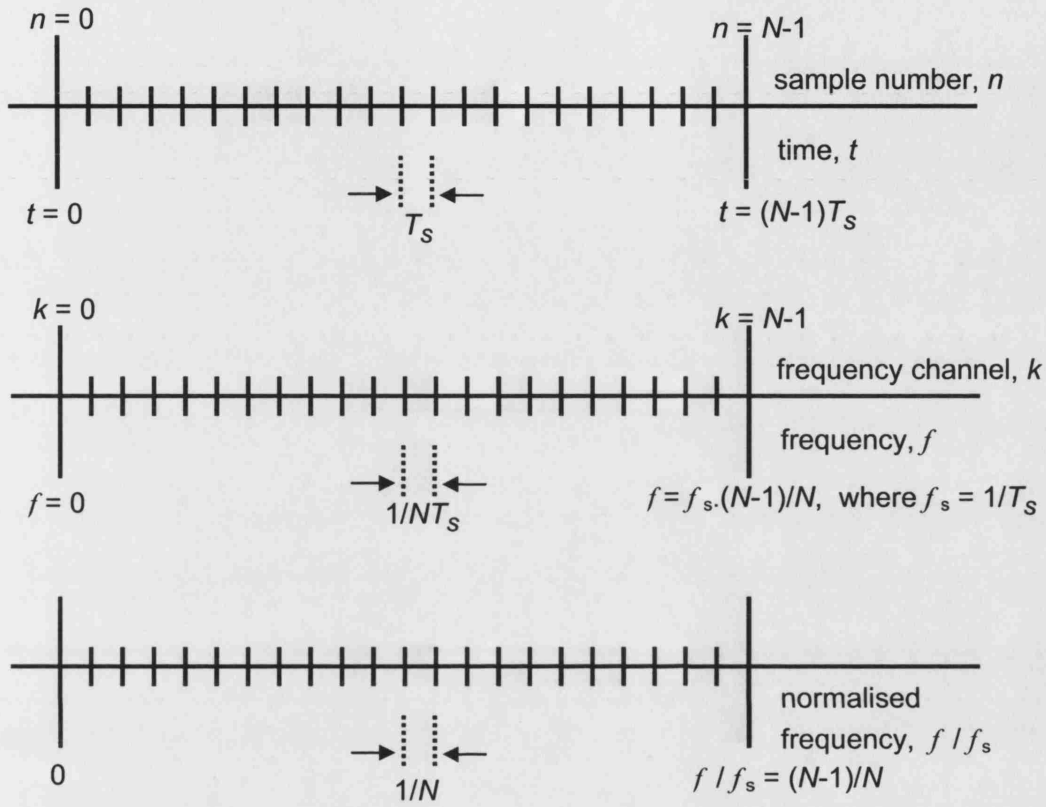

Fig A1 Axes showing integer indexing for the discrete time and frequency domains.

Properties of the DFT: The DFT of a sampled signal $x(n)$, $n = 0$ to N is:

$$X(k) = \sum_{n=0}^{N-1} x(n) \times e^{-j2\pi k \frac{n}{N}}$$

It has the following properties:

- The DFT is periodic such that $X(k) = X(N+k)$, see Figure A2. Normally the range $0 \le k \le N-1$ is used but it is equally valid to use the range $-N \le k \le -1$.
- Part **A** of the DFT in Figure A2 is the complex conjugate mirror image of either section labelled **B** because
  - (i) $\angle X(k) = -\angle X(N-k)$ and $|X(k)| = |X(N-k)|$
  - (ii) $\angle X(k) = -\angle X(-k)$ and $|X(k)| = |X(-k)|$
- For power spectrum estimation only part **A** is required because the power spectrum only requires the magnitude and is invariant to phase;

- For signal reconstruction, or for calculation of the autocovariance function, both **A** and **B** are needed because signal reconstruction requires the phase information.

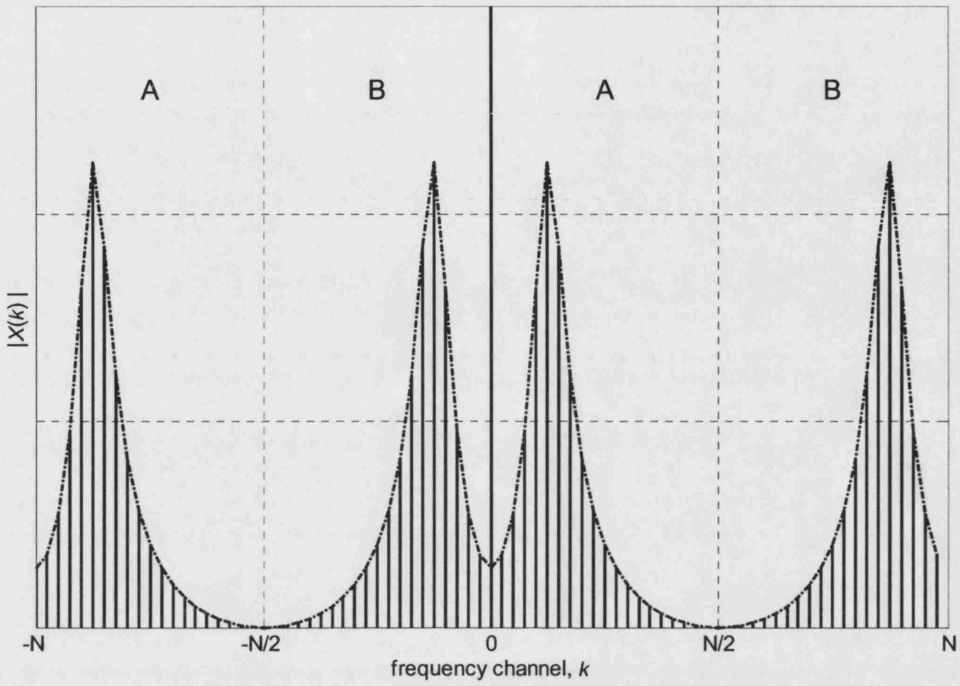

Fig A2 Periodicity in the DFT.

A1.2 Spectral leakage in the DFT

Spectral leakage: Spectral leakage is an effect caused by the finite length of the data sequence which has N samples and duration $(N-1)T_s$. The sequence is treated as though it consists of samples from an infinite sequence which has been multiplied by a window function $w(n)$ with the following properties:

$$w(n) = \begin{cases} 0 & n < 0 \\ 1 & 0 \le n \le N-1 \\ 0 & n \ge N \end{cases}$$

The sampled window $w(n)$ is samples taken from an underlying continuous signal known as a *top-hat* or *rectangular* window of unit height and width T_w. If the window in the time domain contains N samples with sampling interval T_s then $T_w = NT_s$. Note that the interval between the first sample and the last sample would seem to be $T_w = (N-1)T$ because there are $N-1$ intervals between N samples. However, the continuous window is considered to start half a sample before the $n=0$ sample and to end half a sample beyond the $n=N-1$ sample, thus the total elapsed time is $T_w = NT$.

Figure A3 shows a plot of the magnitude of the continuous Fourier transform of the continuous rectangular window for the case when $T_w = 10$. This distinctive shape with side lobes is called a *sampling functio*n. Its value is zero at certain regularly spaced values of f given by:

$$f_k = \frac{k}{T_w}, \qquad k = \pm 1, \pm 2, \pm 3, \ .....$$

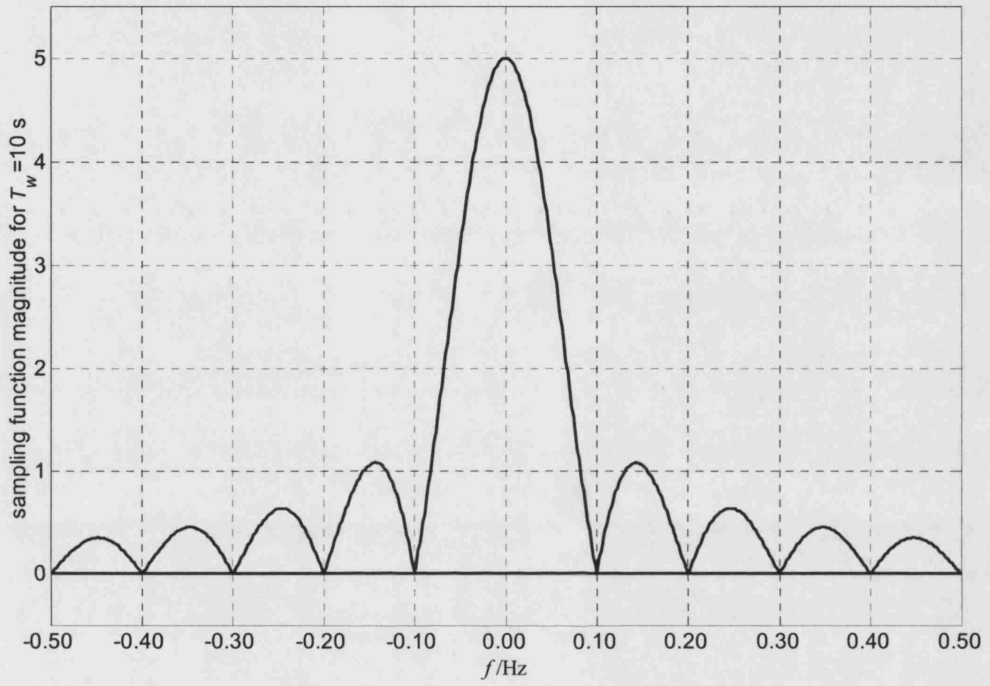

Fig A3 Sampling function of a rectangular window having a width of ten seconds.

DFT of a rectangular window: The DFT is the sampled data version of the continuous Fourier transform. The DFT of N samples from the rectangular window $w(n)$ is:

$$X(k) = \sum_{n=0}^{N-1} 1 \times e^{-j2\pi k \frac{n}{N}}$$

Thus $X(0) = N$ and $X(k) = 0$ for all other values of k. The reason for this is that widths of the frequency channels in the DFT is $\frac{1}{NT_s}$. Since $T_w = NT_s$ the frequencies present in the DFT are

$$f_k = \frac{k}{T_w}$$

These frequencies in the DFT coincide exactly with those frequencies in the continuous Fourier transform where the sampling function is zero. It means that any attempt in MATLAB to visualize the continuous sampling function of Figure A3 by applying the FFT function to $w(n)$

will be unsuccessful because the FFT function does not return any values at the frequencies in between $f_k = k/T_w$ and $f_{k+1} = (k+1)/T_w$.

Spectral leakage in the DFT of a windowed cosine wave: The Fourier transform of a windowed cosine function is the convolution of the sampling function and a spectral peak at the frequency of the cosine wave (and its alias above the Nyquist sampling frequency). The DFT of a windowed cosine wave depends on the number of complete cycles of oscillation captured in the window.

- Case 1: If the window captures an integer number of complete cycles of the cosine wave then the frequency of the spectral peak of the cosine matches exactly with one of the frequency channels in the DFT. For instance, if the sampling interval is $T = 1$ and the *cos* function has a period of 4 samples then a window with 32 samples captures eight complete cycles. The DFT is as shown in the upper panel in the figure below. The sampling function envelopes are shown as a dotted line. They are exactly centred on the $k = 8$ and $k = 24$ frequency channels and there is no spectral leakage because the sampling function has zeros at all the other frequency channels.
- Case 2: If the window does not capture an integer number of complete cycles then the sampling function envelope is centred in between two frequency channels and there is spectral leakage. In the example shown in the lower panel the cosine had a period of 4.27 samples and the window with 32 samples therefore captured 7.5 cycles.

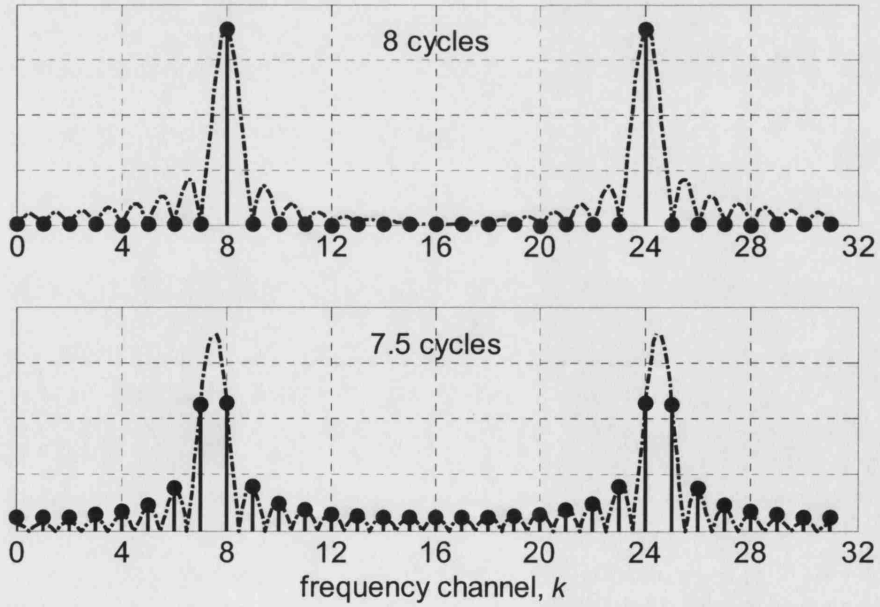

Fig A4 Spectral leakage occurs when a rectangular window does not capture an integer number of cycles of oscillation.

Solutions to spectral leakage: It is clear from the above examples that it is the side-lobes in the sampling function of the rectangular window that are the problem. Other windows have less pronounced side lobes and therefore may be preferred. For instance, the samples in $\{x(n)\}$, $n=0$ to N, may be multiplied by a Hanning window

$$w(n)=0.5-0.5\cos\left(2\pi\frac{n}{N}\right)$$

The Hanning window has a much gentler slope when plotted in the time domain and its transform $W(k)$ has a much reduced side lobes. Further window functions and their specifications and properties are given in Table 9.1. of Mulgrew *et. al* (2002).

The following points can be noted from Mulgrew *et. al* (2002):

- Bandwidth: The frequency domain width of the central peak is smallest for the rectangular window. That is to be expected because the time domain width of the rectangular window is largest.
- Peak side lobe: The side lobes are worst in the rectangular window and better in the Hanning window. A window called the Dolph-Chebyshev window has the smallest side lobes.
- Roll-off: The roll-off is the magnitude of the next side lobe compared to the previous. The Hanning window has the best roll-off.
- Processing loss: Processing loss means that the peak height in the spectrum is mis-represented because the window reduces the overall magnitude of the sampled signal. There is no processing loss for the rectangular window but all other windows have processing loss. The Hanning window has the worst processing loss.

In this thesis only rectangular windows have been applied. The reason for this is that the processing loss would be problematical in a data-driven analysis. For instance, the percentage power calculations in Chapter 4 would not be correct if processing loss were present. In Chapter 6 the problem of spectral leakage when determining surrogate time series was handled in the algorithm by ensuring the data ensemble had an integer number of cycles of oscillation.

A2 ESTIMATION OF THE PERIOD OF OSCILLATION

A2.1 Mathematical formulation

Introduction: This section investigates the estimation of time periods of oscillating signals from the DFT and from zero crossings of the autocovariance function. The author would like to thank Dr Arun K. Tangirala, University of Alberta, for discussions on this topic which prompted the analysis presented here.

The derivations and proof given here are original and lead to this new finding:

- It is more accurate to determine the period of oscillation using the zero crossings in the time domain than by using the positions of peaks in the power spectrum.

The source of error being considered in the DFT is that, for sampled data, the true spectral peak may not be aligned with a frequency channel in the DFT if the data set does not capture a compete number of full cycles of oscillation. Therefore some uncertainty exists in determination of the true frequency and the true period of oscillation. For the *ACF*, the main source of error is that the true zero crossing positions might be in between two lags but is only reported to the nearest integer value. The aim is to study the contributions of these effects to the estimation of the period of an oscillation.

In analysis of data from chemical processes it is the period of oscillation, T_p, that is wanted. The heart of the matter is that Δf errors in determining the exact position of a spectral peak convert to large ΔT_p errors in estimate of the period especially at low frequency:

$$T_p = \frac{1}{f} \qquad \Rightarrow \qquad \Delta T_p = -\frac{1}{f^2}\Delta f$$

Specification of the signal: It is assumed that the signal whose period is to be ascertained is a continuous oscillating signal of the form:

$$\sin\left(2\pi f_p \times t\right) = \sin\left(\frac{2\pi}{T_p} \times t\right)$$

where f_p is the frequency and T_p is the period of oscillation to be determined. The information available comprises N samples from the signal taken at sampling interval T. Thus the samples are:

$$y(n) = \sin\left(\frac{2\pi}{T_p} \times nT\right), \quad n = 0 \text{ to } N-1$$

The period T_p can be expressed as $T_p = p \times T$ where p is not assumed to be an integer. Thus:

$$y(n)=\sin\left(\frac{2\pi}{p\times T}\times nT\right)=\sin\left(\frac{2\pi}{p}\times n\right)$$

It will be convenient later to write $p=p_i+\Delta p$ where p_i is an integer and $-0.5<\Delta p\le 0.5$.

The N samples may or may not capture a complete number of cycles of oscillation. The number of cycles captured is:

$$C=\frac{N}{p}$$

and since C is not necessarily an integer it can be written as $C=C_i+\Delta C$ where C_i is an integer and $-0.5<\Delta C\le 0.5$. It is always possible to find an N such that $C=C_i$ provided p is a rational number.

DFT of the sampled signal: The DFT of the sampled signal is:

$$\begin{aligned}Y(k)&=\sum_{n=0}^{N-1}y(n)\times e^{-j2\pi k\frac{n}{N}}\\&=\sum_{n=0}^{N-1}\sin\left(\frac{2\pi}{p}\times n\right)\times e^{-j2\pi k\frac{n}{N}}\\&=\frac{1}{2j}\sum_{n=0}^{N-1}\left(e^{j\frac{2\pi}{p}n}-e^{-j\frac{2\pi}{p}n}\right)\times e^{-j2\pi k\frac{n}{N}}\end{aligned}$$

Continuing the analysis:

$$\begin{aligned}Y(k)&=\frac{1}{2j}\sum_{n=0}^{N-1}\left(e^{-j2\pi n\left(\frac{k}{N}-\frac{1}{p}\right)}-e^{-j2\pi n\left(\frac{k}{N}+\frac{1}{p}\right)}\right)\\&=\frac{1}{2j}\left(\frac{1-e^{-j2\pi N\left(\frac{k}{N}-\frac{1}{p}\right)}}{1-e^{-j2\pi\left(\frac{k}{N}-\frac{1}{p}\right)}}-\frac{1-e^{-j2\pi N\left(\frac{k}{N}+\frac{1}{p}\right)}}{1-e^{-j2\pi\left(\frac{k}{N}+\frac{1}{p}\right)}}\right)\\&=\frac{1}{2j}\left(\frac{\sin\left(\pi N\left(\frac{k}{N}-\frac{1}{p}\right)\right)}{\sin\left(\pi\left(\frac{k}{N}-\frac{1}{p}\right)\right)}e^{-j\pi(N-1)\left(\frac{k}{N}-\frac{1}{p}\right)}-\frac{\sin\left(\pi N\left(\frac{k}{N}+\frac{1}{p}\right)\right)}{\sin\left(\pi\left(\frac{k}{N}+\frac{1}{p}\right)\right)}e^{-j\pi(N-1)\left(\frac{k}{N}+\frac{1}{p}\right)}\right)\end{aligned}$$

Properties of the DFT: The above function has maxima when

$$\frac{k}{N}-\frac{1}{p}=0\qquad\text{and}\qquad\frac{k}{N}+\frac{1}{p}=1$$

Of these two solutions the $\frac{k}{N}-\frac{1}{p}=0$ solution is a spectral peak below the Nyquist sampling frequency of two samples per cycle (i.e. $f_s/2$) and $\frac{k}{N}+\frac{1}{p}=1$ is an aliased spectral peak above the Nyquist frequency. The second term that contributes the aliased peak will not be considered further in order to simplify the discussion.

The $\frac{\sin\left(\pi N\left(\frac{k}{N}-\frac{1}{p}\right)\right)}{\sin\left(\pi\left(\frac{k}{N}-\frac{1}{p}\right)\right)}$ term is zero when $\frac{k}{N}-\frac{1}{p}=\frac{m}{N}$, $m=\pm1,\ \pm2,\ \pm3,\ldots..$

Positions of the spectral peaks: When N/p is an integer then there are exact integer solutions to $\frac{k}{N}-\frac{1}{p}=0$ and to $\frac{k}{N}-\frac{1}{p}=\frac{m}{N}$ where m is a non-zero integer. The quantity N/p is an integer when the data ensemble captures an exact number of cycles, i.e. when $N=C_i\times p$. In that case the spectral peak is located in frequency channel $k_i=\frac{C_i\times p}{p}=C_i$ and the frequency content in all other frequency channels $k\neq k_i$ is zero.

When N/p is not an integer then there is no exact integer solution to $\frac{k}{N}-\frac{1}{p}=0$. However, k has to be an integer because the DFT is a sampled spectrum. Let $k_p=k_i+\Delta k$ where k_p is the true solution to the equation, k_i is the nearest integer to k_p and $-0.5<\Delta k\leq 0.5$. Then:

$$k_p=k_i+\Delta k=\frac{N}{p}$$

It means that the maximum amplitude of the spectrum is located between two frequency channels of the DFT. Therefore the best available estimate of the true frequency of the spectral peak is k_i and the estimation error is Δk.

The circumstance in which N/p is not an integer arises when the data ensemble does not capture a complete number of cycles of the oscillation. Thus $N=(C_i+\Delta C)\times p$ and so:

$$k_i+\Delta k=\frac{(C_i+\Delta C)\times p}{p}=C_i+\Delta C$$

The conclusions is that the error in estimation of the position of the spectral peak depends solely on the value of ΔC. There is no error when a complete number of cycles is present in the data ensemble, it is worst when the number of cycles present includes an odd half-cycle, i.e. when $\Delta C=0.5$.

A2.2 Estimation of oscillation period from a spectral peak

Estimation of the period of oscillation: The true period of oscillation expressed in samples per cycle is p and may not be an integer. For the case when the data ensemble included a complete number of cycles of oscillation C_i then p can be determined exactly with no error from the peak in frequency channel $k = k_i$:

$$\frac{k_i}{N} - \frac{1}{p} = 0 \qquad \Rightarrow \qquad p = \frac{N}{k_i} = \frac{N}{C_i}$$

The final equality above follows because it was established earlier that $k_i = C_i$.

When the data ensemble is not a complete number of cycles then:

$$k_p = k_i + \Delta k = \frac{N}{p}$$

The true value of p is $p = \dfrac{N}{k_i + \Delta k}$ and the estimate made from the spectral peak in channel k_i is $\hat{p} = \dfrac{N}{k_i} = \dfrac{N}{C_i}$. Making use of the relationship $k_i + \Delta k = C_i + \Delta C$ established earlier the error $\Delta p = p - \hat{p}$ is derived as follows:

$$p = \frac{N}{k_i + \Delta k} = \frac{N}{k_i\left(1 + \dfrac{\Delta k}{k_i}\right)} = \hat{p} \times \frac{1}{\left(1 + \dfrac{\Delta k}{k_i}\right)}$$

$$\Rightarrow \qquad \hat{p} = p \times \left(1 + \frac{\Delta k}{k_i}\right)$$

$$\text{and} \quad \Delta p = p - \hat{p} = -p \times \frac{\Delta k}{k_i} = -p \times \frac{\Delta C}{C_i}$$

The result for Δp shows

- The error in estimation of the period from a spectral peak is zero if the data ensemble captures a complete number of cycles. The worst case error is when the ensemble contains an odd half cycle such that $\Delta C = 0.5$.
- The error is worse when the number of complete cycles in the data ensemble is small which typically happens when the oscillation has a long period.

Comments: The above result indicates that whenever possible a spectral method should use a data set that captures a complete number of full cycles of an oscillation. In practice this aim is not easy to achieve for several reasons:

- The sampling interval cannot generally be adjusted;
- There may be more than one oscillation present. A data set holding a complete number of cycles of one oscillation is not guaranteed to contain a complete number of cycles of another;

- The oscillation may change or die away before the required number of samples have been taken.

The result also shows that the estimation of oscillation period is better when C_i is large, i.e. when large numbers of cycles have been captured. For instance, if 20.5 cycles were present in the data ensemble then the error in estimation of the period would be $p \times \frac{0.5}{20} = p \times 0.025$ or 2 5%.

A2.3 Enhanced estimation of oscillation period from time domain zero crossings

Problem formulation: In a time domain method the period of oscillation is determined from the intervals between zero-crossings of the mean-centred signal. In this section the ideal case of a signal with no noise is considered. It is also assumed that the first zero crossing coincides with the first sample. The practical case where these assumptions are not valid will be discussed later.

Consider the sampling of $y(t) = \sin\left(\frac{2\pi}{T_p} \times t\right)$. The samples are $y(n) = \sin\left(\frac{2\pi}{p} \times n\right)$ where p is the period measured in sampling intervals is denoted by p, and p need not be an integer. The zero crossings occur when $t = m \times \frac{T_p}{2}$. However, the zero crossings coincide with a sample only if p is an integer which is not normally the case.

A integer number C_i of complete cycles of oscillation is considered whose total duration in sampling intervals is ℓ, where ℓ also need not be an integer. For instance, if the oscillation has 10.25 samples per cycle and three cycles are considered then $\ell = 30.75$. Since p is not an integer in this case the zero crossings of the signal do not coincide with the sampling instants and the 3 cycles are competed in between samples 30 and 31.

The quantity ℓ_i is an integer indicating the closest sampling interval to the end of the C_i cycles. An estimator of the period of the oscillation is taken to be $\hat{p} = \frac{\ell_i}{C_i}$. In the above example $\ell_i = 31$, $C_i = 3$ and $\hat{p} = 10.33$ which is close to but not equal to the true value of $p = 10.25$. It is also noted that the $\ell_i = 31$ samples actually occupy more than 3 cycles of oscillation. The exact number of cycles is 3.0244.

Mathematical expressions that summarise the observations in the last paragraph are:

$$p \times C_i = \ell$$

$$\hat{p} = \frac{\ell_i}{C_i}$$

$$\text{with } p = \hat{p} + \Delta p \text{ and } \ell = \ell_i + \Delta \ell$$

Thus:

$$\ell_i + \Delta\ell = C_i \times (\hat{p} + \Delta p)$$
$$\Rightarrow \quad \Delta\ell = C_i \times \Delta p$$
$$\text{and} \quad \Delta p = \frac{\Delta\ell}{C_i}$$

The quantity $\Delta\ell$ is the error caused by rounding the true value of ℓ to the nearest integer and is in the range $-0.5 < \Delta\ell \le 0.5$. Thus $|\Delta p| \le \frac{0.5}{C_i}$.

Comparison of time domain and frequency domain estimates of oscillation period: The time domain and frequency domain errors in the estimate of period are respectively:

$$\Delta p = \frac{\Delta\ell}{C_i} \qquad \text{and} \qquad \Delta p = -p \times \frac{\Delta C}{C_i}$$

where $-0.5 < \Delta\ell \le 0.5$ and $-0.5 < \Delta C \le 0.5$

Both errors are zero when the data ensemble captures a complete number of cycles of oscillation because in that case both ΔC and $\Delta\ell$ are zero. For the worst cases when $\Delta\ell = 0.5$, $\Delta C = 0.5$ the magnitude of the error using the time domain method zero crossings method is *always smaller* since the period p must be greater than or equal to 2 samples per cycle because of the Nyquist sampling theorem. This completes the proof of the original assertion:

- It is more accurate to determine the period of oscillation using the zero crossings of the time domain than by using the positions of peaks in the power spectrum.

A2.4 Worked examples and practical considerations

Worst cases: The worst cases are examined now. As a worked example of the worst case consider a data ensemble that captured 210 samples of a signal of period 10.025 samples per cycle. The exact number of cycles in the set is 10.474, thus $C_i = 10$ and $\Delta C = 0.474$ which is close to the maximum value of 0.5. The period estimated from the spectral peak and its error Δp in $p = \hat{p} + \Delta p$ are:

$$\hat{p} = \frac{N}{C_i} = \frac{210}{10} = 21 \qquad \text{and} \qquad \Delta p = -p \times \frac{\Delta C}{C_i} = -20.05 \times \frac{0.474}{10} = -0.95$$

When using zero crossing in the time domain, then there are 21 zero crossings including the one on the $n = 0$ (the first) sample. They span 10 cycles of oscillation since there are two crossings per cycle plus the first one, thus $C_i = 10$. The closest sample to the 21'st zero crossing is at $\ell_i = 200$ but the true zero crossing of the continuous signal is half way between sample numbers

200 and 201. The error is $\Delta\ell = 0.5$. The period estimated from the zero crossings and its error are given below.

$$\hat{p} = \frac{\ell_i}{C_i} = \frac{200}{10} = 20 \qquad \text{and} \qquad \Delta p = \frac{\Delta\ell}{C_i} = \frac{0.5}{10} = 0.05$$

The error is much higher for the spectral peak method than for the zero crossing method.

A more benign case of the spectral method is when ΔC is minimised by a pre-processing step in which the data set is made as close to an integral number of cycles of oscillation as possible. It could be truncated to 200 samples before embarking on the DFT calculation. If that were done then the 200 samples represent 9.975 cycles of the oscillation and thus $\Delta C = -0.025$. Using the spectral method:

$$\hat{p} = \frac{N}{C_i} = \frac{200}{10} = 20 \qquad \text{and} \qquad \Delta p = -p \times \frac{\Delta C}{C_i} = -20.05 \times \frac{-0.025}{10} = 0.05$$

The pre-processing procedure for truncation requires detection of the zero crossings in order to ensure an integral number of cycles of oscillation so the extra work in calculation of the channel occupied by the spectral peak is not justified since the period can be calculated directly using the already-detected zero crossings using just one division operation.

Noise free signal with arbitrary data ensemble length. The example given earlier used 210 samples from a sine wave of period 20.05. It said that the $n = 0$ sample was a zero crossing and that the 21st zero crossing marked the end of 10 cycles. In a practical case the oscillation may have a phase shift compared to a sine wave. It would often be the case that only 9 complete cycles would be included between the first and last zero crossing even though the data set has more than 10 complete cycles.

Multiple oscillations: If more than one oscillation is present the zero crossings of one oscillation interfere with the pattern of zero crossings of the other. In that case the zero crossing method cannot be used in the time domain.

When noise is present: In a time domain analysis any noise upsets the detection of zero-crossings because it produces additional spurious zero crossings. The problem has been solved earlier in this thesis by working with the zero crossings of the autocovariance function (ACF). The ACF has the same periodicity as the original sine wave but is much smoother. The horizontal axis is a lag axis. It may be computed from the two-sided inverse DFT of the power spectrum, an approach that has the benefit that interfering spectral content such as a second oscillation can be removed from the ACF by frequency domain filtering.

The same considerations apply for oscillation period estimation using zero crossings of the ACF as for the time domain case but the number of complete cycles present is only $C_i/2$ (if C_i is even). Thus the error in estimation of the period is

$$\Delta p = \frac{\Delta \ell}{C_i/2}$$

where C_i is the number of cycles present in the original time-domain data ensemble.

The estimation errors from the frequency domain and ACF are::

$$\Delta p_{ACF} = \frac{\Delta \ell}{C_i/2} \text{ and } \quad |\Delta p_{DFT}| = p\frac{\Delta C}{C_i}$$

where the worst cases are $\Delta \ell = 0.5$ and $\Delta C = 0.5$. The two are equal when $p = 2$

$$\frac{0.5}{C_i/2} = p\frac{0.5}{C_i} \quad \Rightarrow \quad p = 2$$

However, p is always greater than 2 because of the Nyquist sampling criterion so the result shows *the estimation of the period of oscillation using zero crossings of the ACF is always gives smaller worst case errors than the DFT estimates.*

Use of a truncated ACF: In practice a fixed number of cycles of the ACF were used for estimation of the oscillation period. The minimum number of cycles recommended in the ACF was 4. Thus the error in estimation of p is $\Delta p_{ACF} = \frac{\Delta \ell}{4}$. A comparison of the worst case with the worst case DFT method (i.e. when $\Delta \ell = 0.5$ and $\Delta C = 0.5$) gives a cross-over point of:

$$\frac{0.5}{4} = p\frac{0.5}{C_i} \quad \Rightarrow \quad p = \frac{C_i}{4}$$

For instance, if $C_i = 20$ and if one cannot be sure to capture an exact number of cycles in the data ensemble it is better to use the *ACF* method if $p > 5$, while if $p < 5$ the estimates of p will be better using the peaks of the DFT. For practical analysis it is usually preferable if possible to adjust the sampling rate to about 20 samples per cycle (i.e. $p \approx 20$) so in general the *ACF* method gives better estimated of the oscillation period. An example of where the DFT might be useful would be in the case of a periodic but non-sinusoidal waveform having harmonics. With 20 cycles of an oscillation and 20 samples per cycle the DFT method would be useful for estimation of the periods of the fourth and higher harmonics.

A3 DESCRIBING FUNCTION OF A STICTION NON-LINEARITY

A3.1 Overview

Introduction: This appendix presents the derivation of the describing function for a non-linearity having a deadband and stick-slip behaviour. The derivation arose from a collaboration with Shoukat Choudhury at the University of Alberta. As far as Choudhury and the author know, this is a new derivation. The describing function for a deadband is published in many control text books but no prior publication was found of the describing function for a deadband with stiction.

The input and output waveform:

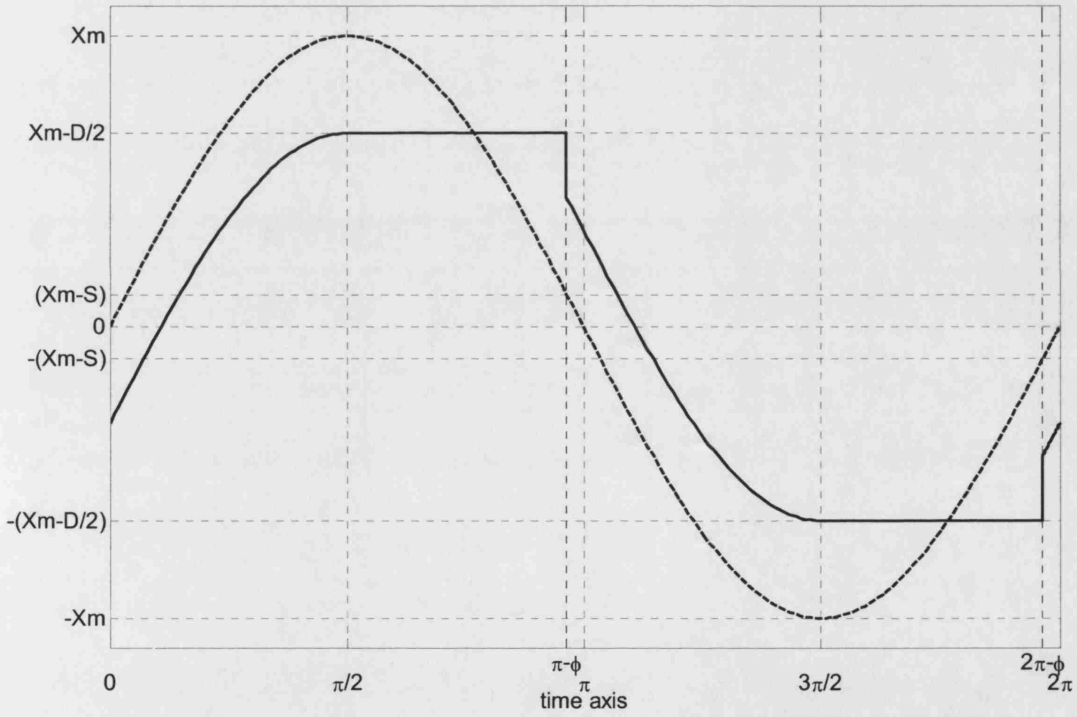

Fig A5 Input and output waveforms showing deadband and stick-slip

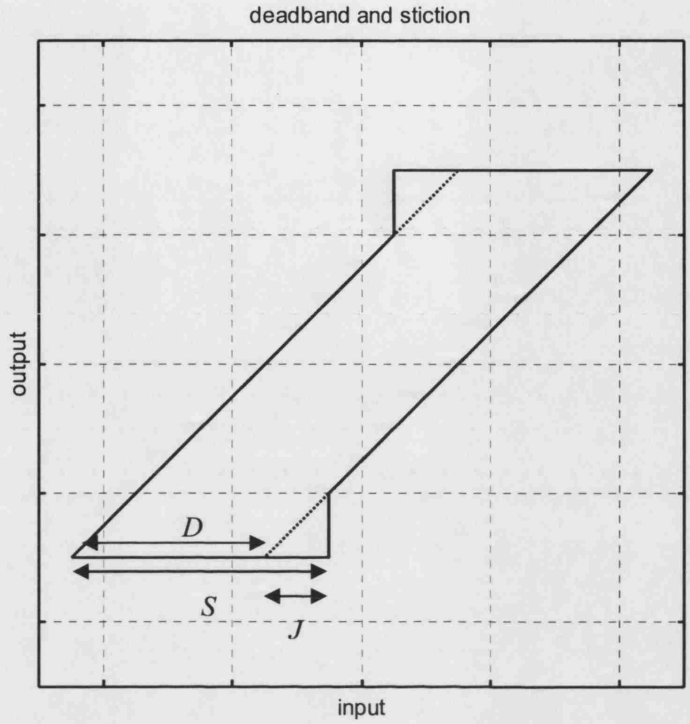

Fig A6 The input-output characteristics of a deadband with stiction non-linearity showing deadband D, stickband J and the sum of the two, $S = D + J$.

Expression for the output of the non-linearity: The output from the stiction non-linearity is not analytic. It is convenient to consider a sine wave input with angular frequency of 1 $\text{rad} \cdot \text{s}^{-1}$ and period 2π as shown in Figure A5. The figure also shows the response from the deadband and stiction non-linearity. The output is piecewise continuous and is described as follows:

$$y(t) = \begin{cases} k\left(X_m \sin(t) - \dfrac{D}{2}\right) & 0 \le t \le \dfrac{\pi}{2} \\ k\left(X_m - \dfrac{D}{2}\right) & \dfrac{\pi}{2} \le t \le \pi - \phi \\ k\left(X_m \sin(t) + \dfrac{D}{2}\right) & \pi - \phi \le t \le \dfrac{3\pi}{2} \\ k\left(-X_m + \dfrac{D}{2}\right) & \dfrac{3\pi}{2} \le t \le 2\pi - \phi \\ k\left(X_m \sin(t) - \dfrac{D}{2}\right) & 2\pi - \phi \le t \le 2\pi \end{cases}$$

where X_m is the amplitude of the input sine wave, D is the deadband, $\phi = \sin^{-1}\left(\dfrac{X_m - S}{X_m}\right)$ and k is the slope of the input-output characteristic in the moving phase ($k = 1$ is assumed for a valve).

A3.2 Derivation of the describing function using Fourier analysis

Evaluation of the fundamental Fourier component: The fundamental component of the complex Fourier series is:

$$\frac{1}{2\pi}\int_{t=0}^{2\pi} y(t)e^{-it}dt$$

where, after substitution of $\sin(t)=\frac{1}{2i}\left(e^{it}-e^{-it}\right)$:

$$\int_{t=0}^{2\pi} y(t)e^{-it}dt = \int_{t=0}^{\pi/2} k\left(\frac{X_m}{2i}\left(e^{it}-e^{-it}\right)-\frac{D}{2}\right)e^{-it}dt + \int_{t=\pi/2}^{\pi-\phi} k\left(X_m-\frac{D}{2}\right)e^{-it}dt$$
$$+ \int_{t=\pi-\phi}^{3\pi/2} k\left(\frac{X_m}{2i}\left(e^{it}-e^{-it}\right)+\frac{D}{2}\right)e^{-it}dt + \int_{t=3\pi/2}^{2\pi-\phi} k\left(-X_m+\frac{D}{2}\right)e^{-it}dt$$
$$+ \int_{t=2\pi-\phi}^{2\pi} k\left(\frac{X_m}{2i}\left(e^{it}-e^{-it}\right)-\frac{D}{2}\right)e^{-it}dt$$

Writing it compactly:

$$\int_{t=0}^{2\pi} y(t)e^{-it}dt = [T1]+[T2]+[T3]+[T4]+[T5]$$

where $[T1]=\int_{t=0}^{\pi/2} k\left(\frac{X_m}{2i}\left(e^{it}-e^{-it}\right)-\frac{D}{2}\right)e^{-it}dt$, and so on.

Evaluation term by term gives:

$$[T1]=\frac{k}{2}(X_m-D)+ik\left(\frac{D}{2}-\frac{X_m\pi}{4}\right)$$

$$[T2]=-k\left(X_m-\frac{D}{2}\right)(1-\sin\phi)-ik\left(X_m-\frac{D}{2}\right)\cos\phi$$

$$[T3]=k\left(\frac{X_m}{4}(1+\cos 2\phi)-\frac{D}{2}(1+\sin\phi)\right)+ik\left(\frac{X_m}{4}\sin 2\phi-\frac{X_m}{2}\left(\frac{\pi}{2}+\phi\right)+\frac{D}{2}\cos\phi\right)$$

$$[T4]=-k\left(X_m-\frac{D}{2}\right)(1-\sin\phi)-ik\left(X_m-\frac{D}{2}\right)\cos\phi$$

$$[T5]=-k\left(\frac{X_m}{4}(1-\cos 2\phi)+\frac{D}{2}\sin\phi\right)-ik\left(\frac{X_m\phi}{2}+\frac{D}{2}-\frac{X_m}{4}\sin 2\phi-\frac{D}{2}\cos\phi\right)$$

Collecting terms gives the wanted fundamental Fourier component of the output:

$$\frac{1}{2\pi}\int_{t=0}^{2\pi} y(t)e^{-it}dt = \frac{1}{2\pi}(B+iA)$$

where $$A = k\frac{X_m}{2}\sin 2\phi - 2kX_m\cos\phi - kX_m\left(\frac{\pi}{2}+\phi\right) + 2kD\cos\phi$$

and

$$B = -3k\frac{X_m}{2} + k\frac{X_m}{2}\cos 2\phi + 2kX_m\sin\phi - 2kD\sin\phi$$

The fundamental component of the complex Fourier series of the input sine wave is $X_m/2i$. Therefore the describing function is:

$$N = \frac{B+iA}{2\pi}\times\frac{2i}{X_m} = -\frac{1}{\pi X_m}(A-iB)$$

A3.3 Limiting cases and other insights

Evaluation of limiting cases: There is no output from the non-linearity when $X_m < S/2$. The limiting cases considered are therefore $X_m = S/2$ and $X_m \gg S$.

When $X_m \gg S$ then $\phi = \sin^{-1}\left(\frac{X_m - S}{X_m}\right) = \frac{\pi}{2}$, $A = -k\pi X_m$, $B = 0$ and thus $N = k$. This result is to be expected because the influence of the stickband and jump are negligible when the input has a large amplitude and the output approximates a sinewave of magnitude kX_m. The slope of the moving phase for a valve with a deadband is $k = 1$ when the input and output to the non-linearity are expressed as a percentage of full range. Therefore for a valve with stiction $N = 1$ when $X_m \gg S$.

When $X_m = S/2$ the result depends upon the magnitude of the stickband, *J*. For the case with no deadband ($S = J$) then $\phi = -\frac{\pi}{2}$, $A = 0$, $B = -4kX_m$ and $N = -ik\frac{4}{\pi} = k\frac{4}{\pi}e^{-i\pi/2}$. For a valve with $k = 1$ then $N = \frac{4}{\pi}e^{-i\pi/2}$. This result describes the situation where the output is a square wave of amplitude X_m lagging the input sine wave by one quarter of a cycle, as shown in the left panel of Figure 7.9.

For intermediate cases where both deadband and slip jump are present then the $X_m = S/2$ limit gives $\phi = -\frac{\pi}{2}$, $A = 0$, $B = -2kJ$ and $N = -ik\frac{2J}{\pi X_m} = k\frac{2J}{\pi X_m}e^{-i\pi/2}$. For instance, if $J = S/2$ and

$k=1$ then the $X_m = S/2$ limit gives $N = \frac{2}{\pi} e^{-i\pi/2}$ and the output is a square wave of amplitude $X_m/2$ lagging the input sine wave by one quarter of a cycle.

When the non-linearity has a deadband only and no stiction ($J=0$) the describing function has a limit given by $N = \varepsilon e^{-i\pi/2}$ where $\varepsilon \to 0$ as $X_m \to S/2$.